GENERATION OF TEXTUAL SUMMARIES AT DIFFERENT TARGET READING LEVELS: SUMMARIZING LINE GRAPHS FOR VISUALLY IMPAIRED USERS

by

Priscilla Santos Moraes

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science.

Summer 2016

GENERATION OF TEXTUAL SUMMARIES AT DIFFERENT TARGET READING LEVELS: SUMMARIZING LINE GRAPHS FOR VISUALLY IMPAIRED USERS

by

Priscilla Santos Moraes

Approved: _____

Kathleen McCoy, Ph.D.
Chair of the Department of Computer and Information Sciences

Approved: _____

Babatunde A. Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____

Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Kathleen McCoy, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Sandra Carberry, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Daniel Leon Chester, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Ehud Reiter, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Vijay Shanker, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

There are many people who have been part of this accomplishment. I am going to start from the very beginning: Mom and dad, thank you! Your example and guidance have been paramount in my journey. You are quite responsible for everything I have achieved. My sister and brother, thank you for the unconditional love and the countless fights; they have definitely made me stronger.

My dear husband, you mean so much to me! I have told you already that this safe harbor you represent in my life makes me believe I can face anything in the world. My kids… All of them! Cindy, my oldest, who was coming when I first started my Ph.D.; Maya, my youngest, who is coming right now as I finish it; and my nieces and nephew, I wish you all had an idea of how much I love you. You have taught me so much! Having you makes me want to be a better person in all senses.

Thank you to my UD family, specially my advisors Dr. Kathy McCoy and Dr. Sandra Carberry, who received me with open hearts and have been so enlightening. Your mentorship has gone above and beyond a Ph.D. advisement and I am eternally grateful for that. Thank you to my thesis committee members for the comments, insights and lessons. Dr. Martin Swany, thank you for bringing my husband (and myself) over to this fantastic institution and for inspiring me to start my Ph.D. program as well, under your advisement.

To all UD professors who have been part of my journey in some way and to the CIS staff, thank you! Friends and lab mates from the Tea House and DAMSL labs, you guys and gals are simply awesome! I wish I could carry you over through my

professional life so we could continue having such enjoyable and nerdy discussions. Thank you to the Cisters organization for inviting me to represent it and to the Women In Engineering (WIE) Steering Committee for giving me the opportunity to serve as a chair and contribute to this outstanding cause of fostering women in STEM fields. To the teachers and staff of the English Language Institute (ELI), thank you for helping me learn this amazing language when I first got to the U.S. eight years ago.

Thank you to the IBM Watson team for having invested in my talent even before I concluded my thesis and for pushing me forward. To my leadership and team at the Watson Labs and Watson CTO – especially Rob High, Ray Chancey and Russell Scott – as well as the Watson Ecosystem, thank you for believing and cheering for me!

Finally, thank you to every single undergraduate and graduate student that I have either taught (and consequently learned from) as a Teaching Assistant/Summer instructor or recruited for my experiments. Same appreciation I feel for the Delaware Association for the Blind and the Harmony School of Sciences from Austin, Texas for the help with recruiting participants for my experiments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This work is concerned with the generation of text at different reading levels by tailoring the generated text to fit the reading level that is appropriate to the reader. The technique is employed in the context of conveying the high-level messages of information graphics in online popular media in order to allow access to such media by people who are blind or visually impaired, as well as by systems with limitations on screen sizes or bandwidth where images are not convenient. The contributions of this work aim to avoid commonly placed rule-based methodologies and to improve different phases of the NLG pipeline.

The methodologies and techniques proposed by this work were employed in the context of the SIGHT system, which provided textual summaries of simple bar charts. In this thesis, we handle single line graphs and have made significant contributions to several modules of the NLG pipeline, including: content determination, text structuring, aggregation and lexicalization.

Texts in popular media are written to target readers at different reading levels – some of the text is rather simple (geared toward 4th grade readers, for example), while other text is quite sophisticated (geared toward college-level readers, for example). We found that text that was geared toward a reading level that did not match the reader was difficult for that reader to understand. Thus, we attempt to produce a summary whose writing sophistication matches that of the article in which the graphic appears. Two of the phases of the NLG system are crucial to achieving generation at different reading levels: aggregation and lexicalization.

The methodologies and techniques developed in the context of this work were evaluated by generating summaries of line graphs present in online popular media. Summaries generated at different reading levels were evaluated by both automatic reading level assessment tools and by readers at different reading levels. The appropriateness of the summary context was evaluated by people with visual impairments who were asked to answer important questions about the presented line graphs.

Chapter 1

INTRODUCTION

Access to online resources and news has grown dramatically in the past decades. At the time this study was first described, Pew Internet & American Life Project Tracking surveys (Center, 2010a, 2010b) showed that around 41% of adults (18 or older) utilized the Internet to get news and have access to popular media resources. Another study showed that various social media websites are the source of acquisition of news by adult users, and that an increasing number of them access these social media sites through mobile devices (Center, 2010a). Assuming that mobile devices might present limitations on displaying images either due to network bandwidth or screen size, having an alternate modality to present the information becomes a desired option. The National Federation of the Blind has estimated the number of blind adult individuals in the U.S to be 25,200,000 (NFB, 2013). From the accessibility perspective, this work has the goal of improving the experience of those blind users that use the internet for access to news from popular media by providing improved and broader access to the content available online.

Information graphics (non-pictorial images such as line graphs, bar and pie charts) are commonly used by authors in order to convey a message or to make a point regarding the topic being discussed; yet screen readers cannot read information in such media and therefore those using them may miss those sources of information. This is problematic because many popular media articles do not repeat the content of their information graphics in the article's text (Carberry, Elzer, & Demir, 2006) and the Alt

text (text that a screen reader can read that an author writes to describe the image) is often missing or misleading. Moreover, technical limitations on either the device (small screen) or the network access (limited bandwidth) reinforces the need for an alternative way of accessing the high-level content conveyed by an information graphic. Based on this scenario, the work presented in (Elzer et al., 2007a) was undertaken in order to recognize the intended message conveyed by simple bar charts. Later, (Demir, Carberry, & McCoy, 2008) developed SIGHT (Summarizing Information GrapHics Textually), a natural language generation system that provides textual summaries of simple bar charts using English language.

Line graphs represent another important way of using graphs to convey messages in online popular media. Due to their continuous nature, these graphs, however, convey a different set of intended messages and pose challenges for identifying and conveying their salient visual features. Therefore, significant additions to SIGHT were required to generate textual summaries of line graphs.

The work described in this dissertation was initiated to extend the SIGHT system to handle line graphs. Many challenges were unveiled during the natural language generation phase where characteristics of the graph and evaluations allowed us to find areas of potential contribution to the NLG field. For example, it was found that line graphs appear in articles intended for widely different audiences, with varying reading levels. We found that in order to be able to generate summaries appropriate for these varying audiences, the text in the summary should be at the same reading level as the article. Therefore, the main goal of this research was to develop a Natural Language Generation system which ***generates textual summaries at different target***

reading levels when describing the high-level content (intended message and outstanding visual features) of line graphs for visually impaired users.

The generation at different reading levels resulted from the desire for the summaries to 1) match the reading level of the text surrounding the graph, since such articles are written for audiences at different reading levels depending on the venue; 2) be understood by the person reading the article. From research and studies performed, we learned that simpler is not always better, and even though previous research has assumed that the simplest text is always the easiest to understand, results from experiments performed showed that the majority of the users at different reading levels prefer text that is tailored to their reading level instead of the simplest text they can get.

In order to address the generation of summaries that are written at different reading levels and that convey the high-level message and important visual features of line graphs, a series of techniques and experiments were performed. These constitute the research contributions of this work and are described in the next section.

1.1 Research Contributions

This work reports on a system which is able to generate descriptive summaries of line graphs at different readability levels. These graphs contain a high-level message that the author is intending to convey. In addition, we would like to convey other visually salient features of the graphic that a reader viewing the graphic might note. We do that by generating a high-level summary of the graphic which delivers the intended message and the graphic's outstanding visual features. The summaries generated vary in readability level, according to the reading level of the text

surrounding the graphic. These two aspects, accessibility and readability sensitive generation of text, comprise the two main contributions this work is concerned.

We developed a set of strategies that affect different phases of the NLG pipeline (shown in Figure 1-1). Specifically, contributions made are related to all of the components of the text planning and micro planning phases with the exception of referring expression generation. The contributions related to Natural Language Generation are:

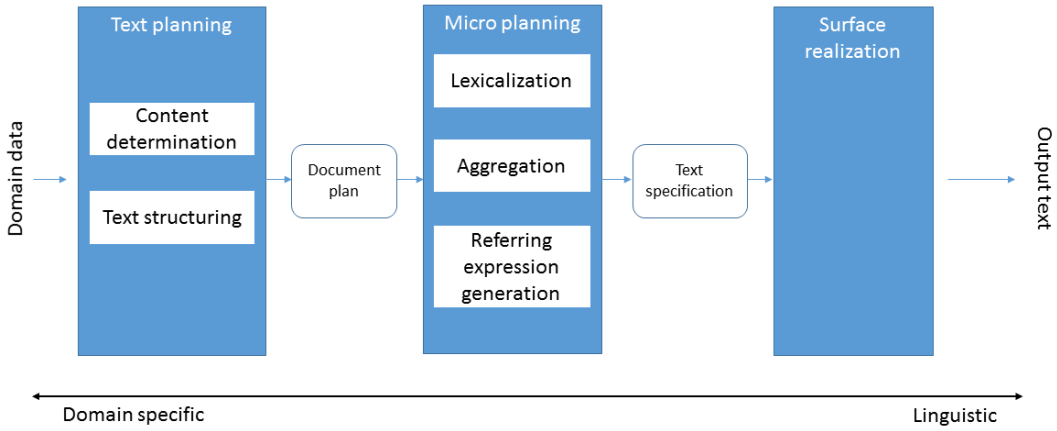


Figure 1-1: NLG system general architecture (Reiter & Dale, 2000).

- Allowing the content determination phase to consider important features of the object being described: The use of a graph-based centrality algorithm for content determination which takes into consideration visual features of the element being described proved to be a great choice. This is enabled by the ability this approach provides to the content determination phase to choose descriptions of the object

which are salient and, at the same time, it implements a discourse-aware technique which produces concise, yet coherent, summaries of line graphs. The content determination phase in this work employs modifications to the PageRank (Xing & Ghorbani, 2004) algorithm in order to consider the presence and intensity of visual features of the line graph being described, allowing the summaries of different graphs to vary both in content and size, depending on the graph itself and which features of the graph are visually salient. Therefore, this approach allows the summary to be customized based on the particular graphic: the more salient features a graphic has, the more detailed is the summary that describes it. This is a promising path to be followed by other NLG systems, independent of the domain the system will be applied to. In contrast to the work described here, the SIGHT system module that generates initial summaries for simple bar charts used logic rules for determining the content to be selected. By using a graph-based approach, the system becomes more flexible and easily adaptable regarding the selection of content for the initial summary.

- Considering additional underlying messages contained in the graphic: In addition to the intended message identified in a graphic, there can be other candidate messages identified in a graphic which augment the high-level message of the graphic and help identify the most important aspects of the graphic. We have developed a strategy that allows this information to help select content for the summary and have tested and validated it. Since these messages carry significant visual information

about the graphic, such information needs to be considered by the content determination module. In contrast, the SIGHT system module that generates summaries for simple bar charts only takes the intended message of the graphic into consideration, leaving candidate messages with non-trivial assigned probabilities out of the scope of the graphic's summary.

- Structuring text in the summary according to the importance of visual features of the element being described: The organization module also applies a graph-based customization technique in which the ordering of propositions in the initial summary is based on the content of the graphic instead of using a predefined organization template. What is deemed to be the most important element present in a graphic will be described first, thus providing more emphasis.
- Allowing the generated text to vary in readability complexity: The micro planning phase aims to generate text at a target reading level. In contrast, most text generation systems are designed for one specific reading level. They usually use measures that will balance the generated text complexity in order to be able to produce understandable text. However, users that access popular media have neither the same age nor the same education and different venues are targeted to users at different reading levels. My hypothesis is that users have preferred online sources which they use to gather information, so if the text that describes the graphic conforms to the overall text in the article, the probability of the graphic summary being understood by the user

increases. In the context of this work, the target reading level of the text to be generated is the same reading level as the one used in the article's text in which the graphic appears. For this, a set of techniques and methodologies were employed and evaluated:

- Determining which syntactic features of a text are associated with different reading levels as identified by readability measurements: From this research, a learning methodology was developed – using decision trees - in order to look at the features of text that can be used while generating text and what their measures are. Learning which values these features take for different reading level text was pivotal to using them during the aggregation phase (part of the micro planning phase as shown in Figure 1-1).
- Enabling aggregation of propositions to be performed efficiently: The system uses a heuristic graph search algorithm to search through possible realizations and aggregation choices. The heuristic attempts to match text feature measures learned from an annotated corpus to achieve the desired reading level.
- Gathering multiple lexical items so that summaries at different reading levels can use different lexical items: An approach for gathering synonyms for the concepts being described was developed. The technique starts with seed words for a concept, captured from a human subjects experiment, that are used to expand the set of lexical items. Synonyms of these seed words

are collected from a thesaurus using only the constraint that both the seed word and the synonym have the same part of speech. Since coming up with the appropriate words to describe a concept is always one of the challenges faced by NLG system architects, the use of seed words from exemplary text provided by humans proved to be a good way of starting from a representative word for defining a concept. The idea behind gathering all of the synonyms that were used in the same part of speech was to allow the lexicalization module to be automated, resulting in a methodology that can be ported to other domains. However, further refinements were required.

- Determining which lexical items are relevant to the domain being described: The set of synonyms gathered from the thesaurus using POS tags was incredibly broad. Many of the lexical items would not be appropriate to the domain under consideration. Clearly, a way of filtering and finding relevant terms was needed. For that, a combination of a language modeling approach and word vectors was used. This combination yielded the best results for filtering the lexical items that were pertinent to the domain of describing line graphs. Our hypothesis is that this technique is robust and scalable enough to be used in different domains as long as the domain concepts can be defined by an initial set of seed words provided to the system.

- Defining the lexicon based on the different target reading levels to increase user understanding of the generated text: After coming up with a set of filtered synonyms for each concept that could be used for describing line graphs, the creation of the lexicon based on grade level was possible by checking for the occurrence of these words in texts marked for the desired grade level in the corpus annotated with the grade level (the same corpus used to learn text complexity related features and their measurements – grade level labeled magazines).

The contributions related to providing access to line graph content to visually impaired users are:

- Extending the SIGHT system to generate initial summaries of line graphs: From the accessibility perspective, the contribution of this work is the development of a natural language generation system for high-level summaries of line graphs. Once the system is paired with a fully robust visual extraction module, it will provide visually impaired users with access to the high-level content of line graphs from popular media (extending the existing tool that gives them access to simple bar charts).
- Considering visual features when building a summary that describes a graph: This allows visually impaired users to have access to an initial summary that goes beyond the graphic's intended message, conveying also its prominent visual features and important candidate intended messages. This approach allows the system to provide a richer,

although still concise, initial summary that covers most of the important high-level information conveyed by the graphic. The initial summary will allow visually impaired users to get the gist of the graphic, with the goal of enabling them to experience the same ability sighted users have when they skim the graphic on a web page.

A minor contribution which allows better usability of the system is described next:

- Enabling users to have access to the system without having to install anything but a web browser plugin: In order to provide SIGHT as a service in the cloud, a Chrome® plugin has been developed to detect an image in a web page and launch the system to generate the summary. This addition now allows the system to be used remotely, since only the plugin must be installed on the user's machine. The system is called using a web socket and the detected image is sent to the server which runs the SIGHT system service. Upon receiving the request, the SIGHT running on the server generates the summary and sends it back to the client.

All of the approaches and methodologies created for addressing the problems found while generating text at different reading levels were evaluated and the results are presented in Chapter 8. Likewise, the usefulness of the initial summaries generated by the system for users who are blind was also assessed and results are presented in the same chapter.

Chapter 2

RELATED WORK ON ACCESSIBILITY OF GRAPHICS

This chapter presents research that has been done in the area of accessibility for visually impaired users. It lists a variety of different research avenues that have been taken in the area of making graphs accessible. The chapter describes various approaches and methodologies which aim to provide access to graphs and charts to people who cannot see them. Different kinds of graphs have been targeted by accessibility research. Some graphs are concerned with showing hierarchical or semantic relationships between entities (with nodes and edges as in a taxonomy, for example). Other graphs, which we call scientific graphs, use bars, pies or lines to show statistical data. Such graphics are used as a device to help the reader interpret data such as data from an experiment. These graphs are important in STEM education and reading and interpreting scientific graphs is an important area of study. A third kind of graph is typically found in less formal situations (such as in popular media). These graphs are placed in a document in order to make a point or to convey a message about one or more entities. Related work presented here ranges throughout all three types of graphics. This work, however, focuses on the third type: graphics that are used in popular multimodal documents to convey a message that augments the message of the document or to provide information on its own.

Some work is related to allowing visually impaired users to access graphs by providing a tactile representation of it. Others provide an audio alternative by employing musical stimuli and some alternatives generate descriptive passages that

narrate the graph. For the different approaches presented here, a comparison is drawn in order to allow the reader to understand how these techniques differ from the objective of this work.

Another relevant area of research which is related to this work is that of Natural Language Generation. Since this aspect of the work represents the majority of the theoretical and practical contributions and implementations, related work in the area will be presented within the respective sections that they have influenced.

2.1 Related Work on Providing Access to Graphs for Visually Impaired Users

2.1.1 The SIGHT system

SIGHT (Carberry et al., 2013; Demir, Oliver, Schwartz, Elzer, Carberry, McCoy, et al., 2010; Elzer, Green, Carberry, Carberry, & McCoy, 2003) is a system designed to generate initial summaries and provide follow up responses about simple bar charts. It is an interactive system where the user navigating the web is provided with the initial summary of a simple bar chart present in the article he/she is reading and is further able to ask for follow up responses, in which case more detailed information about the graphic is selected and translated into sentences. SIGHT is triggered by an add-on for the Internet Explorer® browser (Elzer et al., 2007b) that recognizes the presence of a graphic image (currently differentiating between simple bar charts and line graphs) in multimodal documents from popular media.

Figure 2-1 shows a simple bar chart followed by its generated initial summary. The graphic designer uses a different color for the bar labeled United States. The contention is that she/he does this to catch the attention of the reader and to invoke the

comparison of this bar with all the other ones. Such salient features are very important when the system needs to identify the intended message of the graphic.

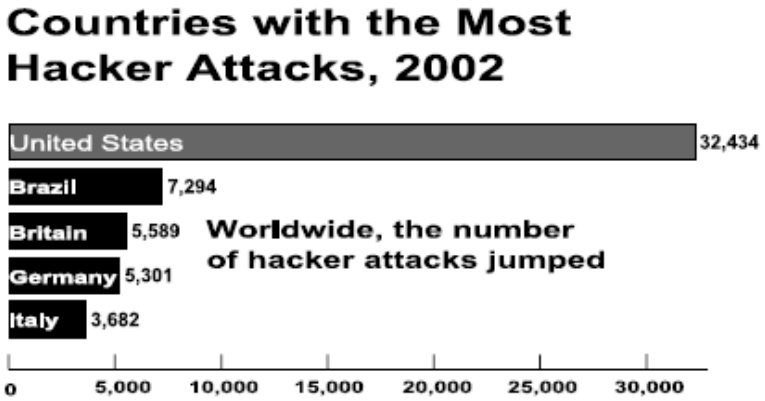


Figure 2-1: Example of a simple bar chart from popular media.

Table 2-1: Initial summary generated by the SIGHT system for the graphic in Figure 2-1.

Initial Summary:
The graphic shows that United States at 32434 has the highest number of hacker attacks among the countries Brazil, Britain, Germany, Italy, and United States. United States has 5.93 times more attacks than the average of the other countries.[1]

The SIGHT system did not handle line graphs, which is one of the contributions of this work. The steps which comprise the pipeline of the SIGHT system are similar for simple bar charts and for single line graphs. However, their implementations differ considerably, particularly regarding the phases of intention recognition, content determination, text organization, aggregation of propositions and

[1] The provided textual summary would presumably be read by a screen reader.

choice of lexical items. Details on the added functionalities to SIGHT in order to handle line graphs are presented in Chapter 3.

2.1.2 Providing access to graphs through audio

A varied set of alternatives for providing access to graphics for visually impaired users has been investigated. One alternative is the use of structured musical stimuli to convey coordinate locations within a graphical grid in order to be able to communicate diagrams composed of geometric shapes (Alty & Rigas, 2005a, 2005b). Sound timbre and pitch ((x,y) and magnitude, respectively) are used to represent the position of the user's cursor. By tracing the object using mappings of timbre and magnitude, size and shape attributes are then communicated to the user through sound. (Brown & Brewster, 2003; McGookin & Brewster, 2006) uses sonification in order to allow visually impaired users to interpret and perform multimodal graph browsing.

The work presented by Kennel (Kennel, 1996) uses an audio-tactile solution to help visually impaired users explore diagrams present in technical reports and papers. In this approach, the graphic is displayed on a touch panel where a part of the graphic can be selected with a finger. An auditory component then presents the selected part to the user by listing the elements on the diagram (describing the element type and its content using words). If a user decides to explore an element further, then an element view and, subsequently, an attribute view can be accessed. The attribute view, which is the most precise display level, tells about the attributes of the element (for frames: position, height, width, shape, area color, shade, line color, line width; for text: position, string, font, font size, font color; for connections: endpoints, nodes, line width, line color, arrows). The authors state that users need to build a mental map of the diagram elements in order to visualize them. They also affirm that congenitally

blind users and users who lost their vision at an early age have more difficulty using the system.

An auditory interface used to display graphics and to help visually impaired users perform steering tasks is described by (Cohen et al., 2006; Cohen & Yu, 2005). The goal is to allow users who cannot see graphs and relational information to navigate through them, moving from one node to another, by using auditory cues to help them with the task. As the diagram is displayed on a tablet, the user can "draw" on the page until the stylus reaches a vertex. When that happens, a sound is played to notify the user about the event. At that point, details about the node are read out loud (for example: the text contained inside the node and a list of its neighbors). The system, called PLUMB, was originally designed to provide access to data structure and relational diagrams to Computer Science majors, but it is also intended to be applied to maps and other representations of such entity-relationship phenomena.

The approaches mentioned would not be applicable to the type of graphs which are the focus of this work. Information graphics cannot be navigated in the same way since they do not reflect vertices and edges representing relationships, but rather they contain high-level messages that convey relational information about a set of entities.

2.1.3 Providing access to graphs through haptic interfaces

Krufka and Barner (Krufka & Barner, 2006) use tactile representations of images in order to make them available for blind users. The image set contained pictures of animals, buildings, people, and objects that were familiar to all subjects that participated in the experiments. This work did not present results on evaluating comprehension of line graphs through use of tactile approaches.

Haptic interfaces are also used to help the user interact with graphics. The work presented in (Ramloll et al., 2000) describes a line graph reader which reproduces the line graph by using audio-haptic displays. The authors map Y coordinates to pitch and include the creation of a sound object that can be positioned by the stylus of the haptic device. Each curve contains its own sound object. The goal of the authors is to increase independence of people with visual impairments when it comes to producing tactile versions of line graphs since the conventional methods were time consuming and complex. With the same intent, other work presented in (Ramloll et al., 2000; Wall & Brewster, 2006; Yu & Brewster, 2002; Yu, Kangas, & Brewster, 2003; Yu, Ramloll, & Brewster, 2001) uses tablets and tactile display systems in order to allow visually impaired users to interact, further investigate and even create bar charts and line graphs.

Even though these initiatives have proven to be successful, their purposes – which generally are to provide access to scientific graphics to help with data interpretation – differ from our purpose which is to provide the user with the high level knowledge conveyed by the graphic. Since the graphics we focus on are used to make a point, providing access to the underlying data does not necessarily achieve that goal since it would require a huge cognitive task. In the types of graphic we focus on, the visual features and the high-level message conveyed are chosen because they make it easier to understand the content of the graphic. From the technology requirement perspective, the initiatives presented in this section have limitations such as the costs associated with the haptic devices required, the fact that some of these systems require preparation work done by sighted individuals, and the experience and knowledge required by people using them.

2.1.4 Providing access to graphs through availability of input data

Another initiative for providing access to graphics for visually impaired users involves the generation of graphics from input data. Goncu and Marriott (Goncu & Marriott, 2008) developed a tool that automatically generates tactile versions of bar and pie charts from input data present in textbooks. The work presented by Yu et al. (Yu et al., 2003) allows blind users to create virtual graphs through the use of a low-cost haptic device. These devices create a graphic from available information and allow the user to explore it through touch.

The GraSSML (Graphical Structure Semantic Markup Languages) approach, proposed by Fredj and Duce (Fredj & Duce, 2007), has the goal of improving the accessibility of diagrams at the creation stage (by making their structural and semantic information available through metadata) as well as making the information "behind" the diagram available for modification and adaptation. The availability of this information is then exploited to generate alternative representations that improve the accessibility of diagrams. It captures the information contained in the diagram and allows the generation of different ways of representing it (through text, graphic, speech, etc.). An example of the Verbalization model is shown in Figure 2-2.

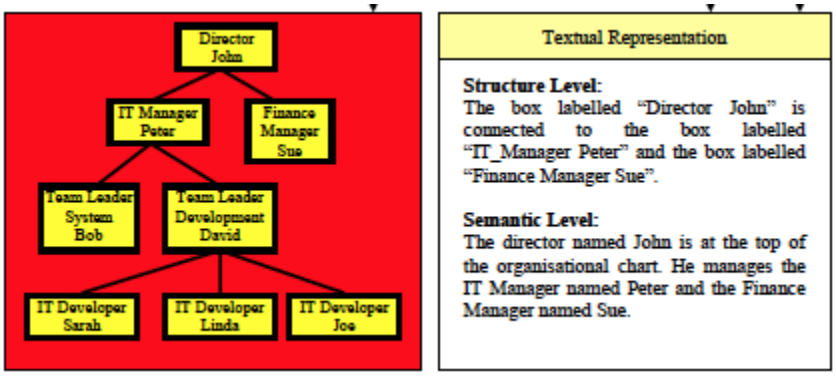


Figure 2-2: Example of the Verbalization model on which a textual description of the diagram is generated by the system described in (Fredj & Duce, 2007)

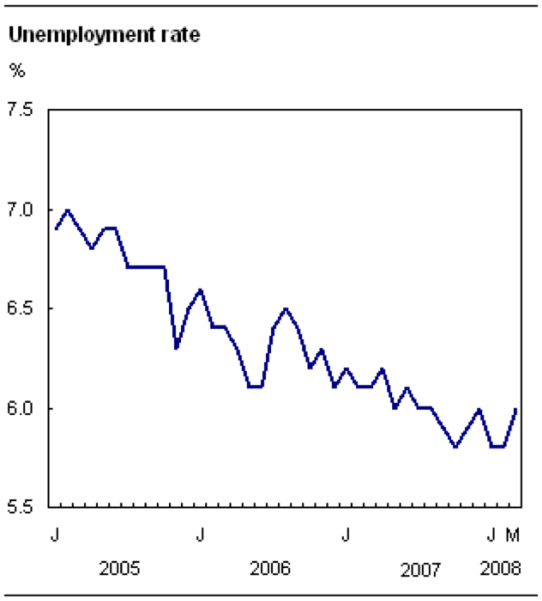
As one can see, these diagrams can have their relationships described, which does not apply to line graphs present in online popular media which are used to make a point. These relationships are provided by human annotators in order to create the metadata and the textual representation then reflects the information on the metadata.

2.1.5 Providing access to graphs through text

(Kurze, 1996) presents the generation of textual summaries of graphics by constructing a description of the diagram by providing its labels, axes ranges, and data set values. The iGraph-Lite system (Ferres, Lindgaard, Sumegi, & Tsuji, 2013; Ferres, Parush, Roberts, & Lindgaard, 2006; Ferres et al., 2007a) provides a template-based summary of what the graph looks like (such as the caption of the graph, and the maximum and minimum values of the graph) based on the data points given in a spreadsheet and allows the user to further explore the graph (Ferres et al., 2007b). An example of a line graph and of a summary generated by the iGraph-Lite system is shown in Figure 2-3. As one can see, these graphs represent statistical data and their purposes are not to convey a message or a point. Therefore, every summary of a line

graph will contain the same information (e.g., max and min values) regardless of what the graph looks like. In contrast, the system described in this thesis will provide different information for different graphics. For example, in the graph shown in Figure 2-3, the falling trend and the volatility of the data points would likely be mentioned. It will provide what it shows to be the most visually salient features of the graphic. The evaluation of the system is presented in (Ferres et al., 2013).

Therefore, the main difference between the approaches described above and our work is the fact that our system automatically captures the high-level message and the outstanding visual features of the graphic and utilizes that information to produce summaries that vary in content and size.



c080404b
This is a line graph. The title of the chart is "Unemployment rate". There are in total 38 categories in the horizontal axis. The vertical axis starts at 5.5 and ends at 7.5, with ticks every 0.5 points. There is only one series in this graph. The vertical axis is %. The units of the horizontal axis are months by year, ranging from January, 2005 to February, 2008. The title of series 1 is "Series1" and it is a line series. The minimum value is 5.8 occuring in February, 2008. The maximum value is 7.0 occuring in February, 2005.

Figure 2-3: Example of a line graph and the summary generated for it by the iGraph-Lite system (Ferres et al., 2007a).

BabyTalk (Gatt et al., 2009) is a system which generates narratives for doctors, nurses and parents from time-series data captured in a neonatal intensive care unit. It is a decision support system designed to extract the most important aspects from neonatal ECG reports. Although the system captures the important times where the results need special attention in order to generate a summary that reflects this phenomena, they do not address the issues related to generating summaries of information graphics the same way we do since their focus is in summarizing key information from a domain/user perspective based on data, instead of describing graphics that try to make a point.

The TREND system (Boyd, 1998a) aims to generate descriptions of time-series data. It uses Wavelet and Fourier Transform in order to segment the trend in a line graph. It is applied to weather data and generates text using FUF/SURGE. It outputs sentences describing each identified trend in chronological order. While this work does describe a line graph, it does not address many of the aspects of line graph description we believe are important. For example, it always uses chronological order (where our system describes the most important things first) and does not include visually salient features in its descriptions. The evaluation shows that 17 of 26 trends that were described by experts were also described by the system. However, the experts also mentioned visual aspects of the graphs such as volatility, max, min, initial and end values of the graphic but the system did not cover such aspects in its generated descriptions. That is one of the main aspects used by our system in order to produce a high-quality summary of a graphic.

2.1.6 Overall comparison

Drawing an overall comparison of our system with other approaches, we can state that: (1) our system intends to provide access to the high-level message and outstanding visual features conveyed by graphics present in multimodal documents from popular media, in contrast with systems that enable the user to explore scientific graphics in detail. Scientific graphics usually do not carry a high-level message but, instead, they require that users have access to the data points in order to draw conclusions and perform analysis. (2) Our work requires neither special equipment nor skills on the part of the user. (3) It also does not require high cognitive load since users do not need to construct a visual representation of the graphic in their minds. The goal of the system described in this work is to make graphics that are already available accessible instead of recreating graphics from input data. Some interviews with visually impaired users indicate that they ignore the graphics in multimodal articles because systems are not available that provide the kind of inexpensive, real-time access that they require. The SIGHT system, to the best of our knowledge, is the only one to consider the high-level message, the point the graphic designer is trying to make through the use of a graph, along with the salient features of the graph in order to create a high-level summary that describes the content of the graph. This provides users with quick access to the high-level knowledge and important information conveyed by the graphics, thereby enabling effective use of this information resource (Carberry et al., 2013; Demir, 2010; Demir, Oliver, Schwartz, Elzer, Carberry, & McCoy, 2010).

2.2 Summary

This chapter presented a variety of accessibility initiatives that aim to provide access to charts and graphs for users who cannot see them. It presented efforts that have been pursued in the audio, tactic and textual areas. It also stated how these approaches and methodologies differ from the work being presented in this thesis. Additional related work, relevant to the area of Natural Language Generation, is described in the specific chapters that they influenced. The next chapter will cover the SIGHT system architecture and its modules.

Chapter 3

THE SIGHT SYSTEM

This chapter introduces the SIGHT system. It describes its overall architecture, its modules and their responsibilities, as well as how they interact to provide the system all the information needed in order to Summarize Information GrapHics Textually.

Line graphs are often used in popular media to convey messages that augment the information provided in the articles in which they appear. These graphics, even with the implementation of the first version of SIGHT, were not accessible to people with visual impairments since the system could not handle such graph type. This work expands the ability of the SIGHT system by modifying several of its modules in order to allow line graphs to be textually described.

The first section discusses the current architecture of the system (Carberry et al., 2013). It outlines the modules responsible for visual extraction of the image containing the graphic (Chester & Elzer, 2005), identification of the high-level message being conveyed by the graphic (Burns, Carberry, & Elzer, 2010; Demir, Carberry, & Elzer, 2007; Elzer et al., 2005; Wu, Carberry, Elzer, & Chester, 2010), and generation of textual summaries (Demir, Carberry, & McCoy, 2012; Moraes, McCoy, & Carberry, 2014a; Moraes, Sina, McCoy, & Carberry, 2014). All of these phases also receive help from additional modules which are also described. The second section highlights the added functionalities needed to handle line graphs in the context of SIGHT. The recognition of the high-level messages conveyed by line

graphs is explained in the third section. This message, along with a detailed representation of the graphic provided by the visual extraction module, are the input to the generation module developed in this work.

3.1 Overall Architecture of the SIGHT System

The SIGHT system consists of five modules. Figure 3-1 shows the architecture of the system, its modules and sub modules. The User Interface module is constructed as a Browser Helper Object (BHO) (Elzer et al., 2007b) in the first version of SIGHT. This add-on for Internet Explorer® browsers is installed on the user's machine, along with the SIGHT system itself, and is responsible for identifying the presence of the graphic image on the Web page being visited. If a graphic is present in the article, the BHO triggers the Interaction Module (IM). The Interaction Module then captures the image file and sends it to the Visual Extraction Module (VEM) where an XML representation of all the graphics features is created.

Based on the graph type (e.g., bar chart, line graph), the XML file is processed by the appropriate Intention Recognition Module where a Bayesian Network is run in order to identify the intended message conveyed by the graphic. The XML is then augmented with the intended message and passed back to the Interaction Module, which then triggers the Generation Module (GM) in order to construct the textual summary of the graphic. The Generation Module is subdivided into four sub modules which are responsible for selecting relevant content (Content Determination), ordering sentences (Text Organization), aggregating sentences and choosing lexical items (Text Complexity), and finally realizing them using the realizers FUF/SURGE (Elhadad & Robin, 1998) for simple bar charts and *simpleNLG* (Gatt & Reiter, 2009) for line graphs (Summary Generation).

The SIGHT system was previously only capable of providing access to the content of simple bar charts and generating textual summaries and responses for this type of graphic. This work extends SIGHT to handle simple line graphs while allowing the summaries to be generated at different target reading levels – such functionality is explained in detail in Chapter 7. Line graphs differ from bar charts in the way they represent data and in the messages they intend to convey. Simple bar charts, for example, might display discrete values at certain points in time regarding one single entity or might convey a single attribute value for different entities. On the other hand, line graphs are usually a time-series representation (or at least an ordinal series) where - sometimes abrupt - changes or steadiness in a trend are displayed, along with some possible peaks and drops in the dependent axis (referred to as the *measurement axis*, as it reflects the observed measurements of the entities being described, in this context). This leads to different kinds of intended messages. Bar charts might be used to convey a ranking of entities according to some attribute or to convey a trend (usually short) over discrete points, whereas a single line graph might be used to convey a long continuous trend or allow one to precisely pinpoint the progression of a sudden drop in the value of an attribute. Visual features are usually different in these two types of graphics. While simple bar charts allow the graphic designer to highlight particular bars (e.g. using a different color), the author of a line graph commonly relies on visual features of the data itself (e.g., sharp changes or annotations at points in the line graph) to draw the reader's attention to interesting intervals.

The steps taken within the system vary based on the type of graphic being processed, especially the identification of the high-level message conveyed by the

graphic. Since our work is focused on generating initial summaries for line graphs, details specific to this phase applied to line graphs are described next.

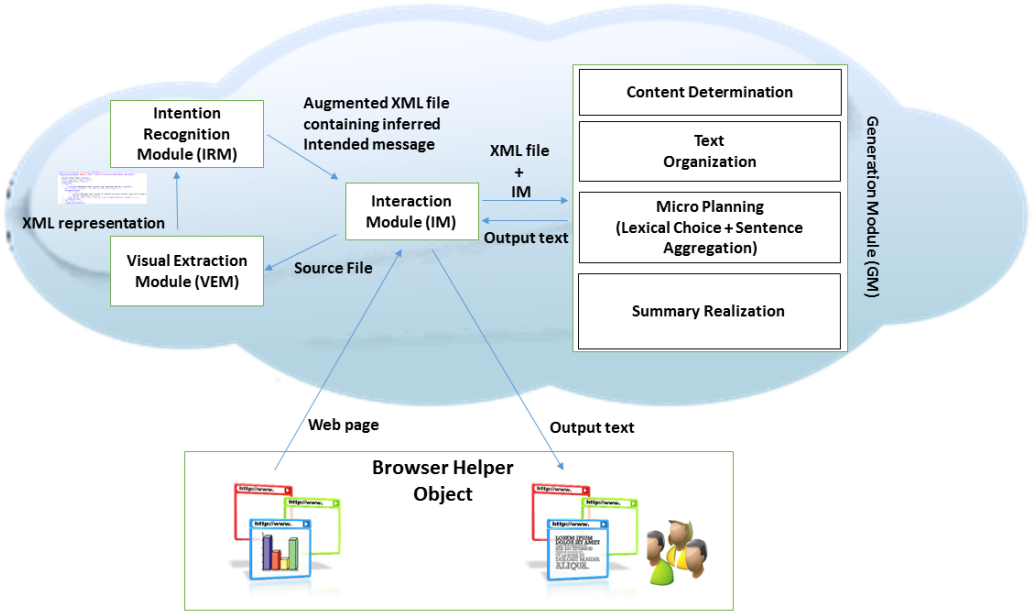


Figure 3-1: Architecture of the SIGHT system (Moraes, Sina, et al., 2014).

3.2 SIGHT System for Line Graphs: Added Functionalities

Enabling the SIGHT system to generate summaries of line graphs required specific modifications to most of the phases of its pipeline. While we concentrate on those involving the Generation Module, other important changes made by other members of the SIGHT group are also mentioned. These include changes to the Visual Extraction Module (Chester & Elzer, 2005) and to the Intention Recognition Module (Wu, Carberry, & Elzer, 2010; Wu, Carberry, Elzer, et al., 2010)

While the previous version of SIGHT required it be installed on the user's computer, this version of SIGHT adds a functionality which allows the system to be used through a web service. A plugin for Chrome®, presented in (Moraes, Sina, et al., 2014), is available and it can be installed on the user's machine. This plugin perceives the presence of an image in the webpage and sends the URI (Universal Resource Identifier) of the image to a server where SIGHT is running, which triggers the whole process from extracting visual features to generating a textual summary.

3.2.1 Changes in the Visual Extraction and Intention Recognition Modules

SIGHT team members modified the Visual Extraction Module to recognize line graphs and be able to generate an XML representation containing sample points (x and y coordinates), labels and tick marks on the axes.

Other team members modified the Intention Recognition Module by implementing a preprocessing step which segments the graphic into visually distinguishable trends. (Wu, Carberry, & Elzer, 2010) presents the work in trend segmentation to help identify the intended message a graphic might be conveying. This step employs a graph segmentation module that, given the representation of the line as a set of small line segments from the visual extraction module, segments the line graph into visually distinguishable trends. For example, the line graph in Figure 3-2 would be divided into two segments, a steady one from 1900 to 1930 and a second rising segment from 1930 to 2003.

As with the original SIGHT for bar charts, after the Visual Extraction Module creates an XML version of the graphic that captures all of the information present in the graphic image, some additional processing takes place in order to enrich the logical representation of the graphic. For example, the XML is preprocessed by the

Caption Tagging Module (CTM) sub module in order to extract clues from text present in the graph's caption which helps the next phase of the architecture to recognize the authors' intended message conveyed by the graphic.

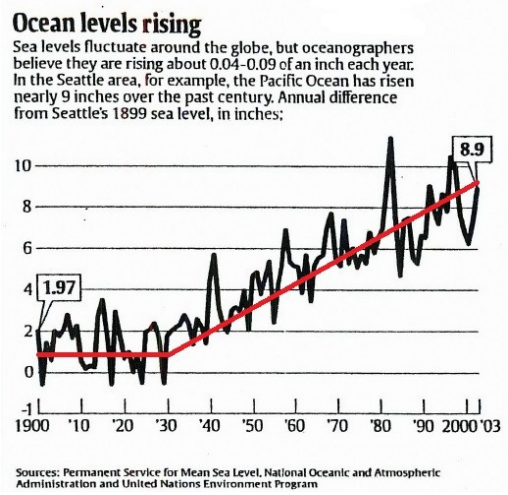


Figure 3-2: Example of a graph segmentation by the Intention Recognition Module (Wu, Carberry, & Elzer, 2010)

Other team members were responsible for recognizing the intended message of line graphs (Wu, Carberry, Elzer, et al., 2010). The set of intended messages identified for line graphs are quite different from the ones used to classify simple bar charts. The possible messages for line graphs identified in the work presented in (Wu, Carberry, Elzer, et al., 2010) are:

1. Rising Trend (RT)
2. Falling Trend (FT)
3. Stable Trend (ST)
4. Big Fall (BF)

5. Big Jump (BJ)
6. Change Trend (CHT)
7. Change Trend and Return (CTR)
8. Change Trend on the Last Segment (CTLS)
9. Contrasting Segment Change Trend (CSCT)
10. Point Correlation (PC)

The Intention Recognition Module (IRM) then uses these segments to suggest candidate intended messages. One such message is created for each category that could be identified from the segmentation produced by the graph segmentation module. For example, candidate messages for the graph in Figure 3-2 would be **Change Trend (stable, 1900, 1930, rising, 1930, 2003), Stable Trend (1900, 1930), Rising Trend (1930, 2003)**. The system will decide which of these is the most likely overall intended message by considering evidence in the graphic.

Information extracted from the graphic by the Caption Tagging Module is one category of evidence. This sub module is responsible for parsing the caption of the graphic to acquire communicative clues from parts of speech that is evidence used to automatically identify the intended message of the graphic designer. For line graphs, verbs such as “change”, “rise”, “jump”, “fall” are considered as communicative signals that can contribute to the classification of the graphic into one of the intended message categories. Relevant nouns and verbs in the caption are therefore considered communicative signals and are taken as evidence that might help in intended message recognition. The XML representation of the graphic is then augmented with the verbs, nouns or adjectives captured by the Caption Tagging Module.

Other evidence includes communicative signals such as annotations on sample points. The identified candidate messages along with the identified evidence are input to a Bayesian Network. The Bayesian Network reasons about the messages and assigns probabilities to them. The message with the highest probability is taken as the intended message of the graphic. The additional identified messages compose the set of candidate messages, and those with non-trivial probabilities assigned to them are also used by the system when selecting content for the initial summary. For example, Figure 3-2 ostensibly conveys that *there is a changing trend in ocean levels: relatively stable from 1900 to 1930 but then rising from 1930 to 2003*. The intended message of the line graph shown in Figure 3-3 also conveys a change in the trend. The message in this graphic is that *there is a changing trend in Durango sales: rising from 1997 to 1999 but then falling through 2006*. Although the Intention Recognition system (Wu, Carberry, Elzer, et al., 2010) classifies both of these graphics into the Changing Trend category, for the graphic in Figure 3-3, it also considers the falling trend to be a strong candidate for the intended message since the end points of this trend are annotated and the caption mentions the action verb "declining". Thus, the Intention Recognition Module of the SIGHT system then assigns a non-trivial probability (~20%) to the Falling Trend category. Such messages will be referred to as *non-trivial candidate messages* throughout this work. As discussed later in Chapter 4, we contend that such messages, with more than trivial probabilities, should influence the summary.

3.2.2 Changes in the Generation Module

Changes in the content determination phase include the consideration of other candidate messages with non-trivial probabilities assigned. While the generation of summaries for simple bar charts did not take into account candidate messages that had

non-trivial probabilities, they are utilized in this work to identify aspects of the graphic that might warrant emphasis in the initial summary; in other words, more emphasis will be given to the falling trend when describing the graphic in Figure 3-3, for example. In contrast, if very low probability had been assigned to this candidate message, the description of the rising and falling trend – which comprise the changing trend – would be given a more balanced description.

Another important concept is that of a visual feature. The presence and intensity of visual features directly affect the content being selected for the initial summary (explained in detail in Chapter 5). Every visually outstanding component of the graphic is considered as a visual feature in the context of this work. These features either are present due to the characteristics of the data, or are added by the graphic designer in order to convey a message. For line graphs in our corpus, fluctuation on sample points (as shown in Figure 3-2) and sharpness of slopes (as in the rising trend of the graphic in Figure 3-3) are examples of outstanding visual features of the data itself. Examples of visual features added by the graphic designer are annotations and words that influence the message conveyed (such as the annotations and the use of the word "**Declining**" in the caption of the graphic in Figure 3-3). Both kinds of features will be called *outstanding visual features* in the context of this work.

We argue that the selection of content and organization techniques both need to be particular to each type of graphic given their peculiarities. Therefore, identifying the most important high-level message conveyed and clearly structuring and organizing a textual summary for different types of graphics require specific and individual tasks. For line graphs, the time-series representation component brings up the need for specific ways of explaining the entity behavior. Merely reading the points

of the line graph demands that users visually recreate the graph representation in their heads. Since this exercise can be very uncomfortable and hard to digest, we want to capture the important features and messages presented in the graphic, organize and aggregate them in a way that is understandable and clear, and realize the knowledge conveyed as concise and coherent natural language text.

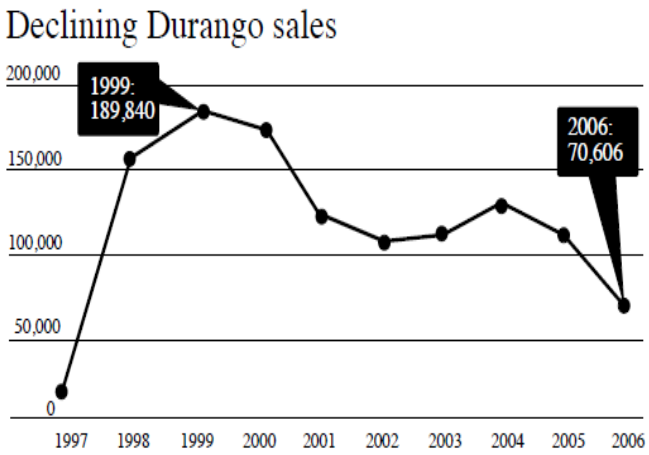


Figure 3-3: Example of a line graph that carries a message of a change in the trend with emphasis being given to the falling part.

Regarding the generation of text, this work proposes an extension to the current generation module of the SIGHT system with significant additions to it. For the micro planning phase, the ability to generate text at different reading levels is introduced. Our hypothesis is that if a user usually reads articles from a given magazine or newspaper, then she or he is comfortable with its reading comprehension level. Hence, generating text that describes a graphic that is part of the article using the same reading comprehension level assures that the text will not abruptly change to a too complex or too simple passage within the context of the article.

Two phases within the micro planning phase are affected when generating text at different reading levels: aggregation and lexical choice. For the aggregation phase, the system measures the reading level of the article's text in order to guide its aggregation decisions. Aspects such as average length of noun and verb phrases, presence of relative clauses, adjectives, adverbs, passive voice, and others are used to guide the amount of aggregation performed and the grammatical construction of the sentences. The frequency occurrences of such features at a particular reading level are learned using a decision tree algorithm. Values of features are then used to build a heuristic for a graph search algorithm, which efficiently searches through possible realizations to find one at the desired reading level.

The lexical choice phase is also guided by the target reading level. The choice of words to compose the sentences of the summary are selected based on the desired reading level from a reading level based lexicon containing concept synonyms by reading level. This lexicon was constructed through a synonym identification process coupled with a reading level filtering step that relies on a corpus annotated with different reading levels. For each concept to be described, possible concept synonyms are identified by a concept expansion phase. The concept expansion is followed by a novel word sense disambiguation phase to increase the quality of the built lexicon by removing lexical items that would not be appropriate to describe concepts in the context of line graphs. This is followed by the reading level filtering step. More details on the micro planning phase (aggregation and lexical choice) are provided in Chapter 7.

3.3 Summary

This chapter described the architecture of the SIGHT system, its modules and their functionalities, as well as the recognition of intended messages conveyed by line graphs. It delineated how the modules of the system communicate and interact. Special description was provided regarding the intention recognition of line graphs since this is the main input to the generation module.

The identification of the intended message and other candidate messages with non-trivial probabilities assigned to them on basis of evidence found in the graphic was introduced. These messages along with salient visual features are used by the Content Determination Module to generate the summaries. This allows the content determination to arbitrate between different aspects of the graphic that can be included in the initial summary and enables the system to create summaries that vary in content and size.

The next chapter provides details on how outstanding visual features of line graphs are identified, extracted and measured to allow summaries to faithfully describe such graphics.

Chapter 4

LINE GRAPH: IDENTIFICATION OF VISUAL FEATURES

This chapter covers the identification of salient visual features present in line graphs that should be included in the generated summary. This affects the content determination phase as it guides the attribution of importance to the various aspects of a line graph. Since different line graphs contain different intended messages and salient visual features, identifying their importance is crucial for allowing the content determination phase to determine what should be included in the summary of the graph.

The chapter describes a human subject experiment performed to determine features that should be considered salient. The experiment design and data collection was performed by another group in previous work in the context of SIGHT. This work uses the data collected and performs analysis of the results in order to assess the importance of visual features in line graphs. This chapter further explains the two identified groups of features and how they affect the content determination phase of the Natural Language Generation pipeline, an object of this work. This chapter also explains how some features of line graphs are measured in order to differentiate their intensity from one graph to another (an important aspect of determining salience).

Line graphs are often used to demonstrate the continuous behavior of an entity (either over time or in relation to an ordinal factor). In order to generate text that describes an entity, it is essential to first identify the features that might be used to describe it. Knowing the importance of a feature amongst the set of all possible

features allows one to choose what is relevant to say and to generate a clear, concise, yet coherent description of the object being described.

4.1 Human Subject Experiment for Identification of Line Graph Important Features

The important features of line graphs that should be included in a summary, or at least what is noticed by humans when viewing them in an article, were identified through analysis of an experiment that aimed to determine what are the features that are included in summaries of line graphs (Greenbacker, Carberry, & McCoy, 2011)[2].

The experiment provided the participants with line graph images followed by a sentence that described the graphic's intended message; the participants were instructed to write a summary which contained what they considered to be the most important information conveyed by the graphic. The goal was to identify people's perception of a graphic, which salient visual features would be mentioned (e.g., steepness of a trend line, fluctuation of data values), and how often they would mention those features. One example of a line graph used in the experiment is shown in Figure 4-1. For this graph the sentence was: *The line graph shows a rising trend from January to December in the number of apple pies sold*. From the experiment we were able to assess features of a line graph that are usually mentioned by users when describing line graphs that contained them.

The experiment used 23 different graphics with various intended messages and visual features. The graphics used for data collection were representative of the set of

---

[2] The experiment was performed and data was gathered by Greenbacker. The analysis of the results and the conclusions reached, which are described in this chapter, are part of the contributions of this thesis.

intended messages and visual features that are common to line graphs (steepness, fluctuation, annotations, etc.). Having a small number of graphs possibly avoided the presence of all the possible combinations of visual features and intended messages the graphs could have. However, since the goal was to identify how often such features were mentioned rather than how they were mentioned when another one was present, the co-occurrence of features was not a crucial fact for the analysis of the collected data. The graphs were extracted from articles of online or paper magazines such as The Atlantic, BBC, Business Insider, Business Week, CNN Money, Forbes, New York Times, The News Journal, Newsweek, USA Today, among others.

Results revealed that 10 main features of line graphs are frequently mentioned by users when summarizing a line graph that contains the feature (with sufficient intensity). The features that we identified from the subjects' descriptions comprise the set of individual propositions that can be used to describe a graphic instance. We use excerpts from actual answers of participants in the experiment to illustrate how those features were usually mentioned:

- The graphic representation form: "the graph shows...", or "the line graph shows...".
- The entity being described by the graphic: "the number of apple pies sold this year".
- The overall behavior of the graphic: "the graph is constantly increasing in stock price...". Since the intended message was provided to the subjects, these excerpts were usually present when the participant meant to further specify it.
- The individual trends forming the line graph: "this graph shows a decrease..., then an increase...".

- The initial/end value/date of the trend/graphic: "From July 28th to August 20th the price of crude oil ranged from about $43 to about $49 a barrel".
- The volatility of the trend/graphic (when sufficiently high): "the crude oil prices has had a very up and down trend, constantly fluctuating".
- The slope of the trend (when it reaches certain values): "the rise from May to August is very steep and the fall from August to November is just as steep".
- The overall amount of change in the trend/graphic: "overall, the stock price decreases from around 39 dollars to around 24 dollars".
- The overall time which the trend/graphic spans: "the x axis goes from January to December...".
- The maximum and minimum points of the graphic: "the sales of apple pies were at their highest in the beginning of December and at their lowest in the middle of January...".

From the set of features identified from the human subjects experiment, we noticed the existence of two main classes: *static visual features* and *valued visual features*. Static visual features are ones where variations in their values do not affect their chances of being included in the initial summary. Initial and end dates, for example, will not have their importance leveraged based on when these dates occurred. While the importance of including static visual features is not affected by their values, their importance might be enhanced by attributes associated with them that are used by a graphic designer in order to draw the reader's attention to that feature. Annotations placed on end points, different bar colors (in the context of bar charts), and bolded legends are examples of visual resources a graphic creator might use in order to give more importance to static features.

In contrast to static visual features, volatility (i.e., the amount of fluctuation) and steepness are two examples of valued visual features in the context of line graphs.

They are more likely to be mentioned when their values are at an extreme, considering the visual presentation of the data in the graphic. This led us to measure the levels of the features to appropriately describe the parts of the graphic that contained them and to influence whether a feature is included in a summary (since it was identified through the subjects' summaries that participants would mention these features in proportion to their intensity). Volatility in a trend or across the whole graphic, for example, varies boundlessly. How extreme the volatility of a line graph is will depend on the variation of behavior of the trends/segments according to their amplitude regarding the dependent axis, for example. The following sections show how we calculate values of valued visual features.

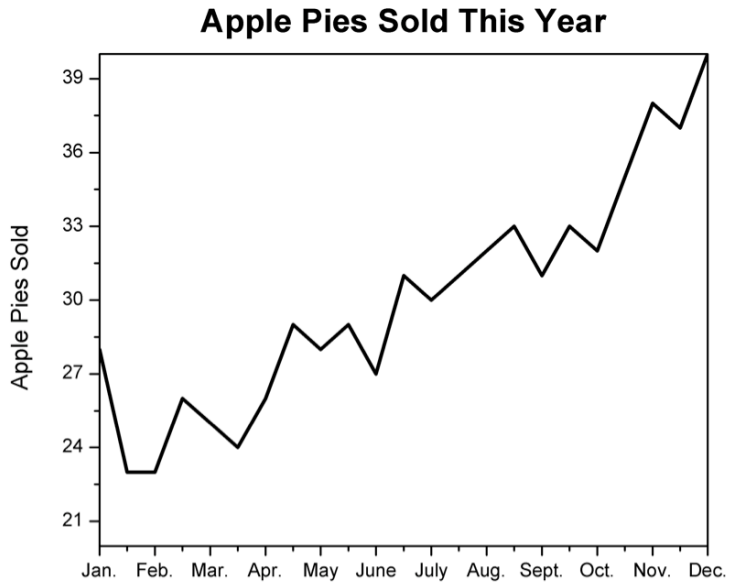


Figure 4-1: Sample of a line graph created for this study.

4.2 Assessing Valued Visual Features

Figure 4-3 and Figure 4-2 show two examples where the degree of volatility differs. It was noticed that, for graphics such as the one presented in Figure 4-2, the participants did not mention volatility (even though the graph is classified as *slightly volatile* by the system), whereas in Figure 4-3 the reference to the high volatility was almost unanimous.

Similarly, the steepness of a trend would likely be mentioned as its slope increased. The graph in Figure 4-2 had its first two trends described as "*sharp increase/decrease*", or "*the trend steeply increases/decreases*" by the subjects quite often, whereas other graphics that presented flat trends or slightly rising/falling ones did not usually include mention of their steepness.

For this reason, two methods were used in order to provide the system with automated tools to deal with these special features. These methods are described in the following two subsections.

4.2.1 Calculating volatility

This section describes the strategy used by our system to measure the level of volatility in a trend. We defined ***volatility*** as the amount of fluctuation in data values that occur in a trend[3].

---

[3] According to (Dictionary.com, 2015), fluctuation as a noun represents *1. continual change from one point or condition to another. 2.wavelike motion; undulation. 3. Genetics. a body variation due to environmental factors and not inherited*. Although the term volatility, also according to (Dictionary.com, 2015), represents *1. a volatile substance, as a gas or solvent*, one of its definition in the Merriam Webster (Merriam, 2016)dictionary is *likely to change in a very sudden or extreme way*. Both of these terms has been interchangeably used by participants to describe the fluctuation of the data points in line graphs.

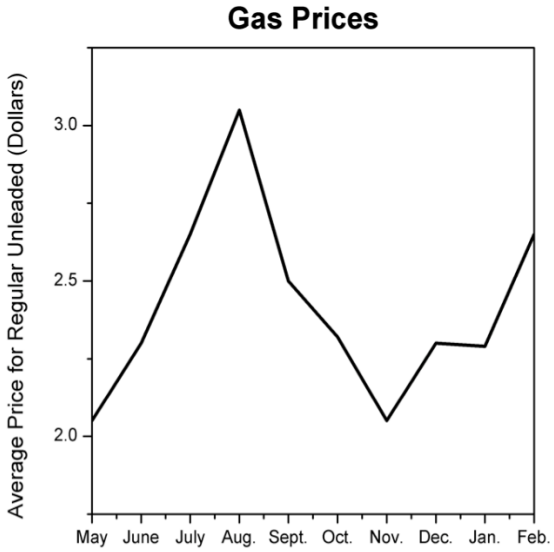


Figure 4-2: Graphic that has its volatility classified as *slightly volatile* by the system.

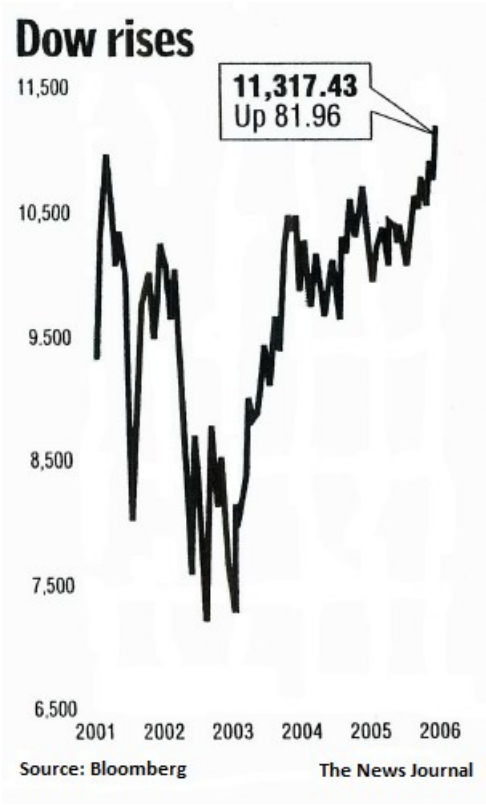


Figure 4-3: Graphic which has its volatility classified as *highly volatile* by the system.

Some work that appears to capture what we want equates fluctuation with coefficient of variation (CV) or relative standard deviation (RSD) (Koopmans, Owen, & Rosenblatt, 1964). However, this is not a good match for us because data values with high standard deviation may not appear to be volatile. For example, when providing the sample points as the distribution for the graphs in Figure 4-2, the CV obtained was 263.06, while the CV obtained from the sample points in the graph shown in Figure 4-3 was 133.34. In contrast, by applying the volatility measurement devised in this work the normalized values for volatility for these two graphs were 0.30 and 0.55, respectively. Therefore, if the trend possesses a substantial number of up-and-down changes and they are not large in their y-value amplitude, the standard deviation will hold a lower value, but our participants often referred to such graphs as highly volatile.

Moreover, it was noticed that users would more likely mention volatility when noticing frequent direction changes in the trend, taking into account the amplitude of those changes as a further factor when defining the level of volatility of a line graph.

We developed a method for estimating volatility that takes into account two aspects. The first aspect is the frequency of change in the trend. This frequency is represented by a change rate, which is obtained by dividing the number of noticeable changes (how many times a trend changed its behavior from falling to rising and vice-versa by at least 3 pixels), herein called fragments, by the number of sample points. The second aspect is the amplitude of those changes. The amplitude is calculated as the ratio of the vertical side of the right triangle formed by the rising/falling trend and the y-axis. To illustrate, let's take a look at Figure 4-4. The bars labeled 1, 2 and 3 represent the amplitude of the fragments used to calculate the overall volatility of the

trend (from point A to point B); while the actual length of the fragment is the size of the fragment (obtained by calculating the hypotenuse of the formed triangle).

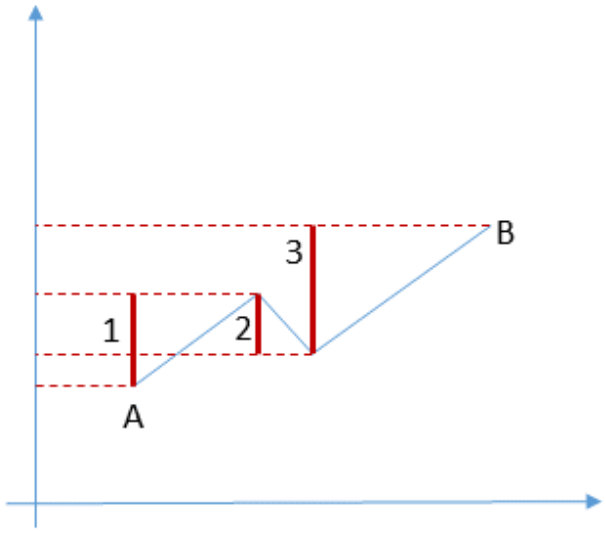


Figure 4-4: Illustration of a trend which contains fluctuation.

Since changes in behavior appeared to be more noticeable (based on the results of the human subject experiment) and associated with the perception of a volatile trend than the amplitude of the graphic trends, a higher weight was assigned to this measure. The formula used to calculate volatility is as follows:

$$volatility = 0.75 * \frac{visual_{changes}}{sample_{points}} + 0.25 \frac{\sum_{i=1}^{f} \frac{size_{verticalSide}i}{yaxis_{amplitude}}}{f} \quad (1)$$

Where *visual_changes* refers to the number of times the behavior (rising or falling) of the sample points changed by more than 3 pixels; *sample_points* is the number of

sample points in the graph or trend (depending on where the volatility calculation is being applied); *size_vertical_side* is the hypotenuse of the triangle formed with the fragment and the y-axis; *yaxis_amplitude* is the length of the projection of the fragment into the y-axis; and *f* is the number of fragments (visually distinguished changes in behavior identified by the Intention Recognition Module) in the trend/graphic (which is 3 in the example shown in Figure 4-4, where three fragments are identified from point A to point B).

The formula used to calculate and classify volatility provides results that are within the range [0, 1] and can be applied to finding the volatility measurement of both individual trends and the whole graph. The results were compared to the way participants mentioned this feature when describing it in order to bucket the values into various description types (described in the next section).

Rising or falling trends that do not present any change in behavior have a value of 0 for its volatility measure. The second part of the formula calculates the averaged amplitude of the fragments that are changing in behavior across the number of trends identified in the graphic in order to assess how large they are in regards to the amplitude of the y-axis.

A volatility measurement rarely gets close to 1 for two reasons. The first is the fact that the number of changes in trend behavior in a graph is unlikely to be the same as the number of sample points in the graphic (obtained from the XML representation of the graphic provided by the Visual Extraction Module described in Chapter 3). Another reason is that the amplitude of all the changes is unlikely to be close to the overall amplitude of the y-axis. Therefore, to obtain volatility measurements that can effectively differentiate among themselves, a normalization step is performed based on

the highest volatility measurement found so far in the corpus[4]. The final volatility value assessed is then used by the system in the content determination step (described in the next chapter) in order to dynamically determine the importance of the propositions representing volatility of various elements in a given line graph.

4.2.2 Calculating steepness

Steepness or flatness associated with rising and falling trends of a graphic is assessed through its arctangent calculation. The final value of the steepness of the trend is represented by the assessed angle and the importance of the feature is directly associated to its slope, as noted in the results of the human subject experiment.

4.3 Describing Valued Visual Features

After values are obtained for both volatility and steepness, another problem needs to be solved: Assuming the importance of volatility and steepness cause it to be included in the summary, how to describe valued visual features based on the resulting value computed for it. Differences in the degree of volatility and steepness need to be described in different ways. This section explains how words are selected in order to describe all the degrees present in these features.

When analyzing the experiment described previously in which subjects provided descriptions of line graphs, not only did the subjects describe the feature itself (the fact that the values were fluctuating), but they also attributed different degrees when describing the feature through the use of adverbs (highly, slightly) or

---

[4] A corpus of 240 line graphs was used for defining the normalization factor. The highest volatility is stored in a configuration file and it can be modified as new graphs are added to the corpus and a new highest is found.

adjectives (many, few). Therefore, besides being able to identify the magnitude of this feature in order to assess the importance of including the feature in the initial summary, the system needs to be able to appropriately describe it using applicable English words.

Based on the result of the calculation of volatility (explained in Section 4.2.1), the system then classifies the trend and the graphic as either: **smooth**, **slightly volatile**, **volatile**, or **highly volatile**. Since the formula used to calculate volatility provides results that are within the range [0, 1], the scale increases at every 33% since smooth is only attributed to trends that have 0 volatility. So if the volatility measured is within the range (0 – 0.33], it is named **slightly volatile**; if it is in the range [0.34 – 0.66], it is named **volatile**; and if it is in the range [0.67 – 1], it is named **highly volatile**. Figure 4-3 shows an example where the system classifies the whole line graph as **highly volatile**, while Figure 4-2 shows a line graph that is classified as **slightly volatile**. These descriptions groups matched the descriptions used by participants in the experiment described in Chapter 4. Matching these descriptions served the purpose of evaluating the calculation and assignment of volatility values to the different volatility intensities.

In order to provide a description of steepness in a trend, an experiment involving human subjects was performed to determine the range of degrees that would be named by participants as: **flat**, **relatively flat but rising/falling**, **slightly rising/falling**, **rising/falling** (no degree identified), **steeply rising/falling**, **very steeply rising/falling**.

The experiment presented trend slopes from every 10 degrees (from 5^{o} until 85^{o} for rising trends and from 275^{o} until 355^{o} for falling trends) in random order and

asked the participants to choose one of the options above to describe the trend. We used Qualtrics (Qualtrics, Provo, UT), a web-based tool for building surveys, for making the questions available to a total of 21 participants. The participants were recruited online and they were from various backgrounds.

We considered a category assigned to a degree range if at least half of the participants had chosen that description. Figure 4-5 shows an example of a trend which has a 45-degree arctangent. Table 4-1 show some results for the experiment. The description for the steepness presented in Figure 4-5 has been chosen by the majority to be "rising" (86% of the participants chose that description). In order for an angle measurement to be classified as slightly steep, steep, or very steep, for example, we looked for an agreement of at least 50% of the participants. When no clear majority was present (as the results for the 55-degree angle, for example), we chose to break the tie on the description side that contains the majority. For this example, since 10 participants chose "Rising", another 10 chose "Steeply rising" and 1 chose "Very steeply rising", the description for this angle was chosen to be "Steeply rising" while there was a tie between "Rising" and "Steeply Rising". The one vote for "Very Steeply Rising" broke the tie in favor of "Steeply Rising" (which is closer to "Very Steeply Rising" than "Rising").

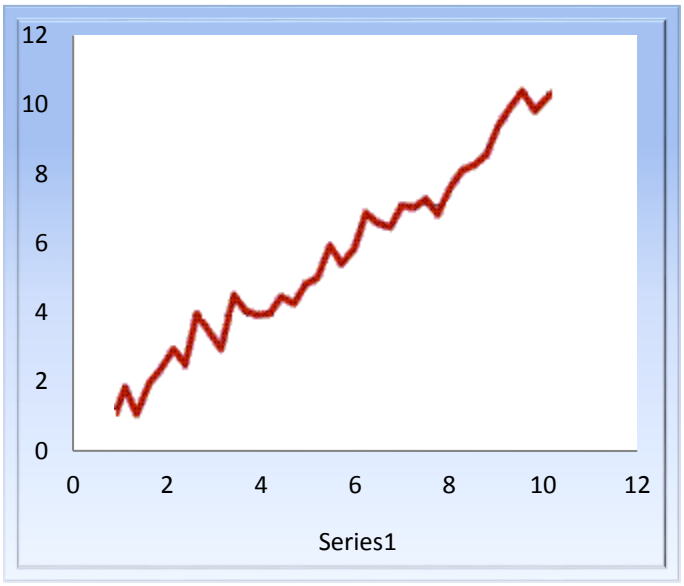


Figure 4-5: Example of a rising trend with a 45-degree incline used in the experiment for assessing how users classified different steepness' descriptions.

Figure 4-2 (page 41) shows a graphic for which the system classifies both trends – the first rising trend and the falling trend - as being **steep** while the second rising trend is classified as **slightly steep** according to the descriptions we gathered from the results of the experiment.

Table 4-1: Example of experiment results for steepness description of a rising trend. (Results were also collected for falling trends).

Degree angle	**Answer**	**Number of participants**	**Percentage of participants**
25	Very steeply rising	0	0%
	Steeply rising	1	5%
	Rising	15	71%

Degree angle	Answer	Number of participants	Percentage of participants
	Slightly rising	5	24%
	Relatively flat but rising	0	0%
	Flat	0	0%
45	Very steeply rising	1	5%
	Steeply rising	2	10%
	Rising	18	86%
	Slightly rising	0	0%
	Relatively flat but rising	0	0%
	Flat	0	0%
55	Very steeply rising	1	5%
	Steeply rising	10	48%
	Rising	10	48%
	Slightly rising	0	0%
	Relatively flat but rising	0	0%
	Flat	0	0%
85	Very steeply rising	19	90%
	Steeply rising	2	10%
	Rising	0	0%
	Slightly rising	0	0%
	Relatively flat but rising	0	0%
	Flat	0	0%

4.4 Describing Static Features Based on Design Choices

The features that describe the overall change in the value (the delta of the y-values over a trend) can be described using either the unit representing it ("the line graph shows an overall increase of 40 million dollars") or by the percentage of the increase. If the y-axis already represents percentage, the generation system will prefer to convey it using the absolute value change (since it will represent the delta already in percentage); thus the importance of the proposition representing the absolute value change is set higher than the proposition representing the percentage of change.

Following the same rule, the features representing the dates can be conveyed by the generation system in different ways depending on the date format of the labels on the x-axis. If the whole graphic or a trend spans over a year or more and the labels present the dates using the format Month/Year, the system will prefer to present this information to the user in months ("the graphic spans over 14 months"). The change value is 36 months, above which the system conveys the time span in years. This is accomplished by boosting the proposition that will best describe a feature given the choices the graphic designer made when creating the line graph.

4.5 Summary

This chapter presented how visual features of line graphs were identified in order to allow the system to generate descriptive summaries of such graphics. An experiment performed with human subjects provided insight on which visual aspects are important to mention when describing a line graph, as well as how important they are by assessing how often they were mentioned. The frequency was found to be directly related to the intensity of valued features (volatility and steepness are two examples).

The chapter further explains and illustrates how valued features are calculated (for the content determination phase) and how we used a human subject experiment to determine which wording is most appropriate for different values of features (for the summary realization phase of the SIGHT pipeline).

The next chapter introduces the content determination module. It describes the role of the features of line graphs - described in this chapter - in the graph-based algorithm, how the features relate to each other, the mechanism for deciding when to stop selecting propositions for the initial summary, and the experiment conducted in order to evaluate the content determination of the system.

Chapter 5

CONTENT DETERMINATION PHASE

The content determination phase of the system is described in this chapter. This phase is crucial since it determines the content of the summaries that are generated for each line graph by choosing which features are important enough to mention in the short initial summary from all possible features that could be said about a graphic.

When talking about a graph, if a trend happens to present attributes such as extreme volatility, high steepness, and annotations on its end points, for example, it is more likely that that trend will be noticed by a user looking at the graphic, so we want to increase the chances that this trend and its attributes are included in the summary of the graph. However, we do not want deterministic rules that will predefine what should be chosen to describe the trend. We want an algorithm that will allow the measurements of importance of a feature (as described in the previous chapter) to make it stand out from the other features when determining the content of the summary. In this example, since the measured values of the attributes of the trend are higher than those of other trends in the graphic, we want this aspect to be considered when determining the content of the summary.

These measurements of importance, combined with the intended message identified in the graph (as well as candidate messages with non-trivial probabilities) and design choices such as annotations on end points, should inform the likelihood of a feature being included in the summary to the content determination algorithm.

Therefore, instead of choosing a threshold to determine which features should be part of the summary, a centrality-based algorithm enables the inclusion in the summary of features that are deemed more important and, along with them, other features that are related to them. These algorithms rely on the popularity of a node and allow the content determination phase of the SIGHT system to employ a discourse-aware methodology when choosing the features to be included in the summary.

Following (Demir, Oliver, Schwartz, Elzer, Carberry, McCoy, et al., 2010), our content determination is based on the PageRank (Page, Brin, Motwani, & Winograd, 1999) algorithm which is intended to determine the importance of a node by how connected that node is to other nodes in a graph. This chapter details the use of this adapted version and the modifications needed in order to accommodate the content determination in the context of line graphs. Key steps are needed in order to use this version of the algorithm:

1. Defining the set of features (herein called propositions): These features are mapped to vertices of the graph and are represented as propositions in the context of SIGHT. The set of identified features is listed in Section 4.1.
2. Setting up the initial importance score of propositions: Propositions are assigned a priori importance based on the assessment of how salient they are in the graphic. The importance of a proposition is also affected by its popularity (interconnectedness), and both aspects contribute to their chances of being selected.

3. Defining the relation between the propositions: Edges of the graph become a representation of the semantic relationship between propositions.
4. Determining the stopping criteria for selecting the most important propositions.

This chapter starts by providing some related work in the area of content determination. It further explains how line graph features and their measurements of importance, assessed by the experiment and formulae described in the previous chapter, affect the content that comprises the summaries. Finally, the chapter presents experiments that evaluate the system's ability to select appropriate content for a summary of a graphic. This evaluation consisted of two phases – where the second phase was the reassessment of the effectiveness of the initial summaries after the system was revised to address the comments and results from the first phase.

5.1 Some Related Work on Content Determination in NLG Systems

The selection of content (a.k.a. content determination phase) is the process by which the information that will be communicated in the text is chosen (Reiter & Dale, 1997). New rules generally need to be created for each domain (Duboue & McKeown, 2003). Different approaches exist for selecting content in natural language generation systems. (Bouayad-Agha, Casamayor, & Wanner, 2011) and (Bouayad-Agha, Casamayor, Wanner, Fernando, & López, 2011) present a content determination approach that is based on the use of a knowledge base to generate football related summaries. The content determination uses inference rules in an ontology to determine which content should be selected and it includes a phase in which the main topic is selected through the use of a user model, a set of weighted heuristics (whose

weights are determined by supervised learning from a corpus of summaries aligned with data) and semantic relations that relate individuals within the knowledge-base.

The work of (Louis, Joshi, & Nenkova, 2010), uses discourse-based relations such as cause, contrast and elaboration, in order to assess text importance for single document summarization. The authors claim that information structure is the most robust indicator of importance, while semantic relations impose constraints on the content determination, having their structure features derived from Rhetorical Structure Theory (W. Mann & S. Thompson, 1987). (Reiter, Sripada, & Robertson, 2003) have investigated ways knowledge acquisition might improve the content determination phase of an NLG system. The authors present experiments using the following Knowledge Acquisition techniques: direct acquisition of knowledge, corpus creation and analysis, structured group discussion, and think-aloud protocol based sessions. As stated by the authors, the first two are the ones widely used in developing NLG systems but the last two were the ones which worked best in their scenario (generation of letters that encouraged smokers to stop smoking) since group discussions allowed the experts to reach an agreement quickly and think-aloud protocols helped with providing good information on reasoning and intentions. They describe how these Knowledge Acquisition techniques can help the content determination phase. (Jordan & Walker, 2005) describe content determination learning for dialogue systems in the context of object description. The authors use previously proposed models in order to define the set of features to be used in a machine learning algorithm that develops a content determination component for generating an object descriptor in dialogue. The utterances are defined based on the set of attributes of an object that are chosen by the object description generator.

The development of the SIGHT generation system faced a challenge due to the non-existence of expert summaries describing the high-level content of graphics. The knowledge acquisition technique that guided our content determination phase was based on direct acquisition of knowledge through judgements by humans who were not necessarily experts on the graphics domains since the graphics vary in their domains of knowledge, and the message we are attempting to convey about the graphics are not domain specific, but composed of the general knowledge they carry. Through the experiment described in the previous chapter we could see that a set of rules would not be enough to depict the varied set of nuances each graphic instance possesses in addition to reflecting how the different combination of feature values could affect each other.

5.2 Setting Up the Initial Importance Score of Propositions

The propositions identified for a particular graphic were defined based on the set of features identified in the experiment described in the previous chapter. These propositions constitute the data pool from which the content determination algorithm will choose the propositions that will comprise the initial summary. Propositions have an initial importance score associated with them. These initial scores are assigned based on four aspects: (1) frequency of the participants' mention of a specific feature in their descriptions of the graphics in the experiment described in the previous chapter; (2) values of valued features calculated by the formulae described in the previous chapter (e.g., the proposition that represents the steepness of a given trend will have a higher initial weight associated with it if it was classified as ***very steep*** than if it was classified as ***slightly steep*** since our experiments showed that the probability of this feature being noticed by a user reading the graphic is higher in the first case.);

(3) enhancements due to graphic design choices such as annotation on the end points; (4) the fact that a proposition belongs to a trend which is part of the intended message or a candidate message. These factors combined indicate the initial importance of a proposition.

Overall, the inclusion of a proposition in the summary of a line graph is not a yes or no decision based on whether or not the measurement of a visual feature exceeds a predefined threshold (Carberry et al., 2013). Ranking of a proposition is affected by its connectivity with other propositions (vertices) in the graph, which represents its popularity, as well as by its a priori importance.

5.2.1 Setting a priori node importance in PageRank

A node's importance score, as computed by the PageRank algorithm, is influenced by the weights associated with the number of edges between that node and other nodes. However, besides the popularity of a node (how much connectivity it possesses) we also want to consider its isolated importance, defined by its initial importance score. The goal is to capture in the algorithm the intrinsic ranking of visual features of line graphs in isolation.

None of the versions of PageRank, however, allows a priori importance associated with the nodes themselves. The PageRank formula inherently relies on the weights on the edges, and ultimately converges to an importance score associated with a node. In the context of SIGHT, features (represented as nodes) were identified as possessing an initial importance that differentiated them from the other features in the graph. In order to capture this importance in a way that complies with the use of the PageRank formula, a special vertex called *priority* was introduced by (Demir, Carberry, & McCoy, 2010) and also applied in the content determination of this work.

This node is never considered for inclusion in the summary, but it is instead connected to every other node in the graph, with the connecting edge weights assuming the values of the feature weights assessed through the human written summaries.

The weights, defined by the four aspects described earlier, are assigned to the edges connecting nodes to the *priority_vertex*. Valued features such as volatility and steepness have their initial weights boosted based on the degree assessed (described in Section 4.2); dates and values that correspond to the format chosen by the designer to present the data are boosted by a predetermined factor when annotations are associated with them; propositions that are part of the intended message and candidate messages with non-trivial probabilities assigned to them are boosted by the messages probability assigned by the Intention Recognition Module.

In this scenario, if a set of propositions captured as nodes are related to both the intended message and another candidate message, these propositions will have their initial weights boosted twice, increasing considerably its likelihood of being chosen as part of the initial summary. To illustrate this functionality, we present an example in Section 5.5 where two graphics (one which also has a candidate message and another one which only has the intended message) have their set of chosen propositions affected by the propagation of their propositions' importance.

5.3 Defining the Relation Between the Propositions

(Demir, Carberry, et al., 2010) presents a framework where semantic relations between propositions are represented by the edges of the graph and it introduces the concepts of **attractors** and **repellers.** These relationship types are responsible for defining how a node is related to its siblings and ancestors/descendants.

Attractor relations relate two propositions that should generally be discussed together, while repellers relate propositions that generally should not be discussed together (e.g., because they would provide redundant information). The relations classified as attractors are: ***complementRelation***, ***belongsToRelation***, and ***contrastRelation***. One example of ***complementRelation*** is between the nodes **trend_initial_date** and **trend_end_date**, since the information about the end date of a trend is usually used to complement the information about the start date of the same trend (the same applies to **trend_initial_value** and **trend_end_value**). The ***belongsToRelation*** connects propositions that comprise parts of a whole to their respective wholes. For example, proposition nodes that describe a trend in the graphic (such as **trend_initial_value**, **trend_volatility**, **trend_steepness**) are related to the **trend_description** proposition (e.g., "rising trend", “stable trend”). The ***contrastRelation*** connects propositions that convey, as the name suggests, contrasting information about a concept such as the **graph_max_point_value** and the **graph_min_point_value** propositions.

The ***redundancyRelation*** is a **repeller** relation type. This relationship connects propositions that are conceptually similar, but might convey the information in a different way. The proposition **trend_overall_value_change** (which tells how much a trend either increased or decreased in its unit) is related to the proposition **trend_overall_rate_change** (which conveys the percent or rate of the change) through the ***redundancyRelation***, since both represent the same concept, only varying in the way the information is represented.

These relations were established based on how two propositions were mentioned together or apart in the human-written summaries provided by the

aforementioned experiment. In order to reflect this using the PageRank algorithm, the edges representing attractor relationships are given high scores, causing two nodes that attract each other to mutually raise each other's scores. In contrast, edges representing repeller relations have a much lower weight so as to propagate a smaller weight between nodes that repel each other. Figure 5-1 shows a model of a graph.

The nodes of type **candidate_message** and **trend_description** can vary in number depending upon the number of candidate messages suggested by the Intention Recognition Module and number of trends present in the graphic itself. One category of intended message is *Change-Trend-Return* that contains three trends: a changing trend followed by a return to the direction of the first trend. In cases where there is more than one candidate message (as in the example provided by Figure 3-3, page 32), there will be a node of type candidate_message to represent each one.

Each relationship type in the graph receives a weight that is responsible for reflecting the relation type behavior between two nodes. These weights will provoke an attraction or repellence between nodes. These relationship types can also be in various degrees if the relationships have different importance within the domain. In our case, for example, ***contrastRelation*** is not as attracting as the ***complementRelation*** since providing complementary information about a feature appeared more frequently in the human written summaries than providing contrasting information. That suggests that the weight associated with ***complementRelation*** should be higher than the one associated with the ***contrastRelation***.

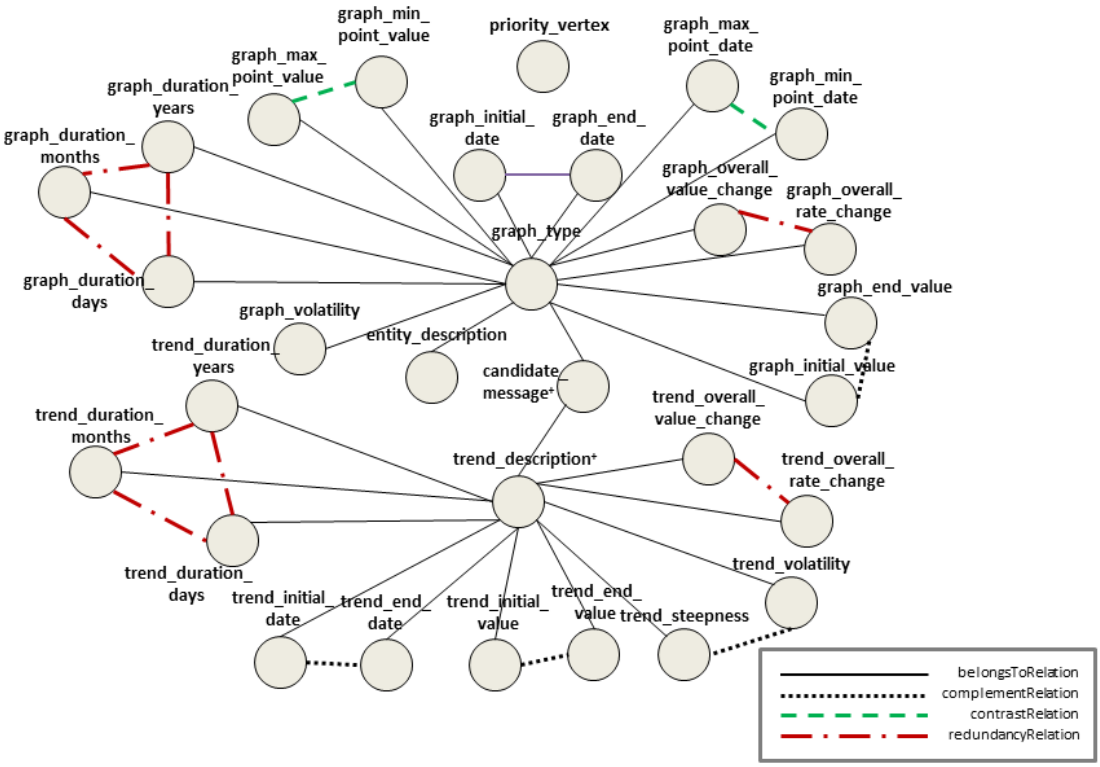


Figure 5-1: Graph nodes and their relationships. belongsToRelation, complementRelation, and contrastRelation are the attractor relations. RedundancyRelation is the repeller relation.

5.3.1 Selecting propositions in a discourse-aware fashion

The most straightforward way to use PageRank for natural language generation would be to run the algorithm one time (until it converges) and then take the top *n* propositions to include in the summary. It could possibly, however, generate either a redundant or an incomplete summary where the *n* most important features present could be conveying the same information (the propositions describing a trend's absolute change value and its percentage of change) or even missing complementary information that is usually mentioned together (as noticed in the experiment described in Chapter 4) such as the minimum and maximum values of the graphic.

To address this issue, (Demir, 2010; Demir, Carberry, et al., 2010) proposes a discourse-aware content determination framework where propositions are selected one at a time and their semantic relationships are considered in order to attract information relevant to the current proposition and repel any information that is redundant to it. This strategy allows the selection of each proposition to be affected by its relationship with the previously selected propositions, thereby increasing the power of the algorithm to select more relevant information. This work also adopts this strategy on its content determination phase.

Recall that the relationships between nodes are of two kinds: attractors and repellers. These relations assign weights to the edges connecting two propositions. Once a proposition is selected, propositions connected to it by attractor relationships have their importance increased by the propagation of a high factor (therefore making them more likely to be selected next), while propositions connected to it by repeller relationships have their importance increased by a much lower factor, therefore making them less likely to be selected next. The steps taken in order to accomplish this are:

(1) Run PageRank to convergence in order to identify the most important proposition and include that proposition in the summary (mark proposition as selected).

(2) Update the weights on the edges that are connected to the selected node by:

a) Raising the weights on attracting edges by a high factor and

b) Raising the weights on repelling edges by a low factor.

(3) Repeat the process starting from running PageRank again to select the next proposition until a stopping criterion is met.

The formula used for this adapted version of PageRank, presented by (Sinha & Mihalcea, 2007), is:

$$PageRank(V_a) = (1-d)+d \sum_{(V_a,V_b)\in E} \frac{w_{ba}}{\sum_{(V_c,V_b)\in E} w_{bc}} PageRank(V_b) \quad (2)$$

where w_{ba} is the weight associated with the edge between vertices (V_a) and (V_b), E is the set of all edges, and d represents the damping factor, which is currently set to 0.15 in this work as opposed to the default value of 0.85 used in the previous version of SIGHT, since the probability of making a random jump in our scenario is very low (it is an undirected graph with no isolated nodes).

By altering the weights between each run of PageRank, the system is more likely to select a set of propositions that attract each other and avoid choosing propositions which repel. An important question this algorithm leaves open is how many propositions to include in a summary. The stopping criteria defined in this work are explained in the next section.

5.4 Determining the Stopping Criteria for Selecting the Most Important Propositions

Deciding what to say in the initial summary using PageRank was only the first step of the content determination process. The decision about when to stop using the highest ranked nodes after each iteration of the algorithm imposed a challenge. (Demir, Carberry, et al., 2010) arbitrarily chose a set number of propositions in a response. This work, on the other hand, uses heuristics based on differences of importance scores from one selection to the next in order to define when to stop selecting propositions for the initial summaries. Initial summaries are intended to be concise and, at the same time, complete enough to give an overall idea of the most

important information contained in the graphic. The relative importance scores of nodes seemed to be a promising boundary determiner, but it still required a definition regarding comparing importance scores in order to define the cut off. In this scenario, it is paramount to be able to tell when a node, given the importance of the previous selected nodes, is much less important, and stop.

Table 5-1: Pseudo code showing the two strategies for the stopping criteria in PageRank.

```
for each node{
   infoScore1 = node.score;
   infoScore2 = node.next.score;
   if absolute difference between infoScores < averageScoreGap
      add node to listOfNodes;
   else
      break;
}
if listOfNodes is missingRequiredNodes{
   listOfNodes.clear
   for each node{
      infoScore1 = node.score;
      infoScore2 = node.next.score;
      if absolute difference between infoScores < averageScoreGap / 2
         add node to listOfNodes;
      else
```

```
            break;
        }
    }
```

Two strategies are used in the stopping criteria (as shown in Table 5-1): The first technique calculates the average gap between importance scores of all nodes in the graph; this average defines the threshold used for determining the selection of new nodes since, empirically, this average split the set of important and less important nodes well for the set of line graphs that were examined. The algorithm keeps selecting nodes until the difference between the current node and the next node to be selected is greater than the calculated average gap. The second one only comes into play when following the stopping criterion above results in a summary that does not contain the minimum required nodes for that graphic instance (graph type, entity description and intended message)[5]. This generally happens when the average gap is too big (as noticed in some cases), for which the algorithm then keeps selecting nodes but now it compares the difference between importance scores with the average gap divided by two. The more trends a graph has, the higher is the number of important features competing to be part of the initial summary.

When the algorithm reaches another stopping point, it then makes sure that the required nodes were included. If any of these propositions are still missing, the algorithm continues including nodes until all of them are part of the initial summary.

---

[5] These three propositions were shown to be essential to the basic understanding of a summary as presented in previous SIGHT work (Demir et al., 2008; Elzer et al., 2008).

The examples shown in Table 5-2 reveal that the number and the type of selected propositions are not the same, showing how our stopping criteria selects the more relevant information customized for each graphic.

5.5 Example of How Features and Candidate Messages Affect Content Determination

The examples in this section, based on the graphics in Figure 5-2 and Figure 5-3, illustrate how the content determination approach using a modified version of PageRank selects a different set of propositions for two graphics that possess a similar overall structure. For this example, one of the graphs present in the corpus was revised in order to modify its salient visual features and candidate messages with non-trivial probabilities assigned by the Intention Recognition Module. Table 5-2 shows the set of propositions selected for the graphics and it also demonstrates how the stopping criteria played a role in determining the amount of information that should be added to the initial summaries for the two different graphics.

The graphic in Figure 5-3 was adapted from the graphic in Figure 5-2 so we could clearly visualize how our version of the PageRank algorithm would select different propositions to include in the initial summary of these graphics. For the adapted version of the graphic, we did not include the sample point annotation of the falling trend and the title no longer contained the word “declining”. The exclusion of these visual features causes the Intention Recognition Module to assign a higher probability to the changing trend category because it no longer assigns much weight to the falling trend alone (higher than the 78% assigned to the original version of the graphic, where the falling trend category receives a probability of 20% since it contains the visual features shown).

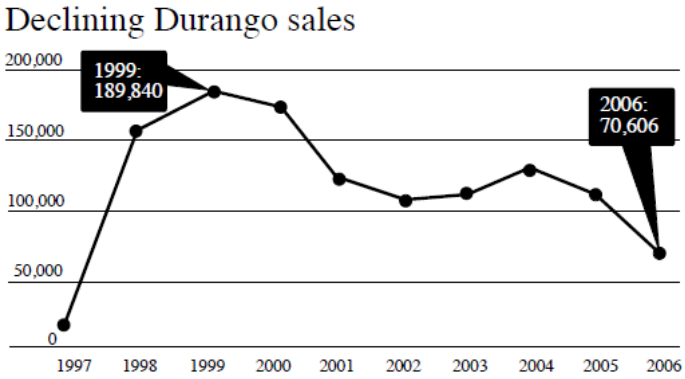


Figure 5-2: Original graphic present in the corpus.

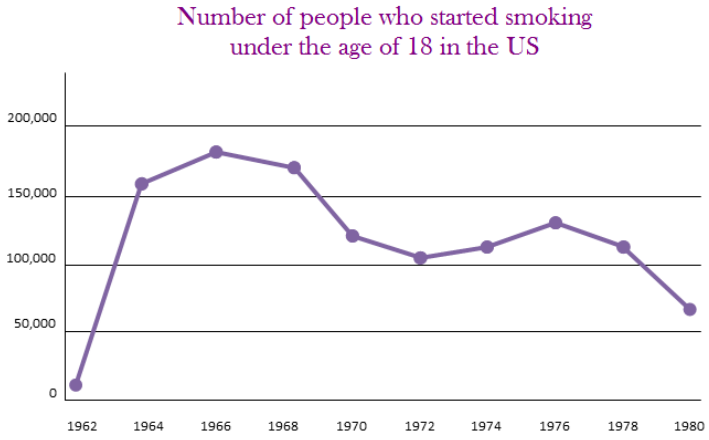


Figure 5-3: Instance of a graphic created for this study.

Since, in the case of the original graphic, the falling trend and its features were boosted by the 20% assigned to the falling trend candidate message, the set of propositions selected by PageRank for the two graphics are different in number and content. We can see that the end points of the falling trend were selected for the initial summary of the original graphic but not for the summary of the adapted graphic. This happens due to both the presence of annotations on the sample points and the falling

trend belonging to a non-trivial candidate message. We can see that in the adapted version the rising trend was selected first. That is due to the fact that, in this graphic, both trends (rising and falling) would probably have the same initial importance. The rising trend, however, was initially boosted by more important propositions connected to it (the steep slope is one example).

Table 5-2: Selected propositions (shown in order of selection) to be included in the initial summary of the example graphics and the summaries generated for each graphic.

Propositions for graphic shown in Figure 5-2 **for which the IRM assigns 78% to changing trend and 20% to the falling trend categories**.	**Propositions for graphic shown in Figure** 5-3 **for which the IRM assigns 97% to changing trend and 1% to the falling trend categories.**
Node: entity_description Description: the number of Durango sales Membership: line graph Node: graph_type Description: line graph Membership: line graph Node: composed_trend Description: CHT Membership: line graph Node: trend_description Description: falling trend Membership: FT Membership: CHT Node: trend_description Description: rising trend Membership: CHT Node: composed_trend Description: FT Membership: line graph	Node: entity_description Description: the number of people who started smoking under the age of 18 in the US Membership: line graph Node: graph_type Description: line graph Membership: line graph Node: composed_trend Description: CHT Membership: line graph Node: trend_description Description: rising trend Membership: CHT Node: trend_description Description: falling trend Membership: CHT Node: steepness Description: 57.0948

Propositions for graphic shown in Figure 5-2 **for which the IRM assigns 78% to changing trend and 20% to the falling trend categories**.	**Propositions for graphic shown in Figure** 5-3 **for which the IRM assigns 97% to changing trend and 1% to the falling trend categories.**
Node: trend_end_value Description: 189840 Detail: annotated Membership: CHT Membership: rising trend Node: trend_initial_value Description: 189840 Detail: annotated Membership: FT Membership: falling trend Node: trend_end_value Description: 70606 Detail: annotated Membership: FT Membership: falling trend	Detail: steep Membership: CHT Membership: rising trend
Summary generated by the system: *The image shows a line graph. The line graph presents the number of Durango sales. The line graph shows a trend that changes. The changing trend consists of a rising trend from 1997 to 1999 followed by a falling trend through 2006. The second segment is the falling trend. The falling trend has an initial value of 189840. The falling trend has an ending value of 70606. The first segment is the rising trend.*	**Summary generated by the system**: *The image shows a line graph. The line graph presents the number of people who started smoking under the age of 18 in the US. The line graph shows a trend that changes. The changing trend consists of a rising trend from 1997 to 1999 followed by a falling trend through 2006. The first segment is the rising trend. The rising trend is steep. The second segment is the falling trend.*

The use of PageRank for content determination in the context of line graphs has required substantial changes such as taking into consideration candidate messages and how they might affect the choice of which pieces of the graphic should be conveyed in a response. Adding this functionality to this current content

determination module and running it on the corpus of line graphs allowed the assessment of how candidate messages particularize a summary. They raise the importance of features that represent pieces of the graphic that are visually outstanding, since these same features led the Intention Recognition Module to also consider that message as the possible intended message of the graphic designer. We claim that this information is too valuable to be left behind when describing a graphic for a person who is unable to see it.

5.6 Evaluation of the Content of an Initial Summary – Phase 1

To evaluate how well the system was selecting the most important content of a graphic for the summaries, and consequently their completeness and conciseness, a pilot human subjects experiment was performed. 16 graduate students from various Computer Science areas were recruited as participants. The participants were provided with the initial summaries generated by the system and were asked to draw a rough sketch of what they believed was in the original graphic. The initial summaries were not aggregated (one sentence was used to realize each proposition) and pronominalization was not performed. The following is an example of a summary used in the experiment generated for the line graph in Figure 5-4:

The image shows a line graph.
The line graph presents the value of Dow.
The line graph shows a trend that changes.
The changing trend consists of a big fall from 2001 to 7/2002 followed by a rising trend through 2006.
The first segment is the big fall.
The second segment is the rising trend.
The rising trend has an ending value of 11317.43.

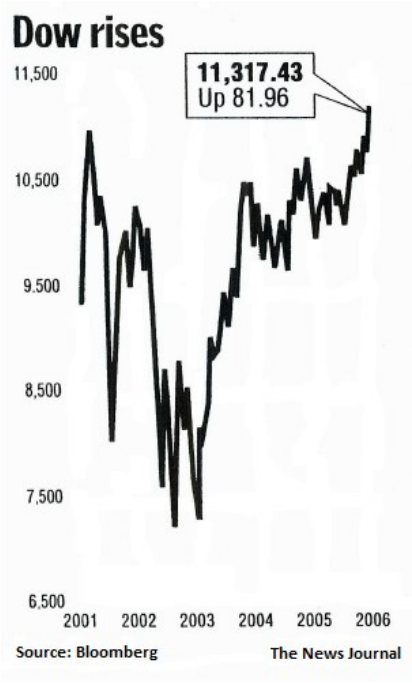


Figure 5-4: Example of a line graph used in the content determination evaluation experiment.

The purpose of asking the participants to draw the graphic was not to see if they could faithfully reproduce the graphic (since the goal of the system is not concerned with being able to draw the graphic from the summary, but with identifying the information content from it). Rather, the purpose of having the subjects draw the graphic was to ensure that they processed the summary in enough detail so they would be able to judge its appropriateness.

After drawing the graphic, the participants were shown the original graphic image and were asked a set of questions to assess how effective they considered the initial summary to be. They were asked to assign a rating (1 being strongly ineffective and 10 being strongly effective) to the initial summary, to point out any misleading or unnecessary information present in the summary, and to describe what they felt was

missing and should be included in the initial summary. Each participant processed between 10 and 15 graphics in a 45-minute period. These graphics covered a variety of recognized intended messages, as well as a variety of outstanding visual features. The set of graphs originally used in the experiment contained 2 graphs (L5 and L17) which overlapped with the graphs in the experiment for feature identification, described in Chapter 4. The results from including the two graphics and not including them are provided next.

The average rating given by all participants to all graphics (including the two graphs used for feature identification) was 7.54, which indicates that the subjects overall found the summaries having an above average effectiveness. When removing the responses about the two graphics used in the experiment for feature identification, the average rating dropped to 7.48. Of importance for evaluating the stopping criteria is information the participants found to be missing or misleading. The following are a subset of the comments provided by the participants:

- The number of subjects affirming that the initial summary was missing relevant information was relatively small. Only 49 out of 201 responses pointed out the need for more information content when describing what they felt was missing in the summary. Some individuals, indeed, declared that the initial summary was missing the initial and end values of some of the trends. However, there was no consensus about this. Upon further reflection, the task that we gave the participants to draw the graphic may have influenced their feeling that they wanted the end values so that they could faithfully draw the graphic. Nonetheless, having the participants draw the graph was

essential to make sure they would pay enough attention to the initial summary and be able to judge if it was carrying the most important information about the graphic or if it was misleading.

- Some participants stated that the initial summary was repetitive. That aspect might have been mentioned for two reasons: the first was the fact that the sentences were realized without being aggregated and the second was the presence of individual trend statements and of initial and end dates propositions in the summary in cases where this information was already provided by the sentence conveying the intended message. The repetition issue was subsequently addressed by both aggregating sentences and pronominalizing referring expressions. The repetition of initial and end points was addressed by adding a repeller relation between initial and end points in the graph and intended message nodes in the PageRank graph. In that case, whenever the intended message was selected, the end points that belonged to trends pertaining to the intended message would be repelled, so they would not be included in the same response.
- Some of the observations made by the participants did not have any relation to the content selected by the system, but it was rather related to either the organization of the sentences or to the repetition of the referring expressions. These results showed how important the aggregation step is in generating effective summaries. They also motivated tackling the aggregation phase for the system as it is discussed in Chapter 7.

Although the overall experiment results were very positive and indicative that the methodology and stopping criteria are reasonable, it was found that the participants had a difficult time evaluating the actual content because text realization decisions got in the way. Thus, the decision to implement some aggregation and then re-run the experiment with a larger number of participants was made, so that a more accurate assessment of the content evaluation could be performed. The second experiment intended to reflect the evaluation of the content determination module without being affected by discourse issues found in the first experiment's results. The results of the second experiment are discussed in the next section.

5.7 Evaluation of the Content of an Initial Summary – Phase 2

The second evaluation of the content of summaries for line graphs produced by the system had the same configuration as the first experiment and, surprisingly, similar results. The participants were provided with graph summaries and the same steps were taken in order to assess the quality of content determination in the summaries. The difference between the two phases of this experiment is in the presentation of the summary. For the first phase the summaries presented one sentence for each proposition selected by PageRank, which contained some redundancy and repetition of referring expressions, while in the second phase summaries had some aggregation and pronominalization applied to the sentences. The following summary contains these modifications for the graph presented in Figure 5-4:

The image shows a line graph, which has many peaks and valleys, presenting the value of Dow. It conveys a trend that changes consisting of a big fall from 2001 to 7/2002 followed by a rising trend through 2006. The rising trend is sharp and has many ups and downs. It has an ending value of 11317.43. This trend follows the big fall, which is also sharp and shows much fluctuation.

With the aggregation plus pronominalization step the system was able to eliminate the dissatisfaction that arose from the repetition and the redundancy present in the previous summaries. However, there was still some dissatisfaction now due to some of the summaries being considered "too complex" by some of the participants. 29 undergraduate students from various majors (Cognitive Science; Chemistry; Communication Interest; English; Physics; Criminal Justice; English; Political Science; Computer Science; Civil, Chemical, Electrical, Mechanical, Environmental Engineering) participated in the second phase. They were able to answer a total of 331 evaluations, an average of around 11 graphs per participant. Even though the results for the second phase were expected to be higher than the ones for the first phase (since aggregation and pronominalization were introduced) the average score assigned by all the participants for the summaries was slightly lower (7.30) when considering all the graphics (including the two graphics which were part of the experiment for feature identification). When the two graphics used for feature identification were removed from the results, the average rating dropped to 7.08.

From the results of this phase of the experiment, it was noticed that the users' taste of appropriate aggregation varies across users. From that observation, the contention that a system should be able to assess a user's preferred level of text complexity and adapt to it was born. It further implies that using the same reading level as the article containing the graph to guide the level of aggregation applied to the summary allows the summary to more appropriately fit the article's text and the user's reading level of preference.

The decision to apply different reading levels also required an adjustment to the vocabulary (the lexicon used in the generation of the summaries). For summaries

which should be at a lower reading level, a set of words more commonly found in text appropriate to that level must be used in order to describe things such as volatility and steepness of a trend (this decision was also inspired by the comments from the experiments in which some subjects wrote that they did not know the meaning of the word *volatility*). The choice of a different lexicon for the aggregation used in the second phase of the experiment increased the vocabulary approval by 25%. However, we still needed to strive to find even more appropriate wording for describing visual features of line graphs since we still received comments from participants about that matter.

Also based on previous results, the system was adjusted in order to eliminate redundancy of text and content. Adjustment based on comments about some information being repeated (by modifying the relationship types between those propositions in the content determination algorithm) and the aggregation plus pronominalization eliminated the repetition caused by reintroduction of entities. It was noticed that the system adjustment was effective since almost no subject complained about repetition (only 2 out of 331 mentioned that we should have aggregated some additional sentences further).

Two other identified points of discussion provided by the participants were the sentence organization and the lack of detailed information about the graphic (initial, end, maximum and minimum values; rate of change). The former gives rise to an interesting problem. Our claim is that the most important things in a graphic are usually noticed first by a sighted user, who pays attention to the outstanding visual information first, and to the complementary features second. It is the system's goal to provide a summary that also focusses on the most important pieces of the graphic first.

The lack of detailed information might be addressed by providing the ability to ask follow-up questions (as described in Chapter 9). It can be considered that this information is not crucial to the initial summary since the goal of the initial summary is to convey the intended message and the outstanding visual features of the graph (things that a sighted user catches with just a gist of it). One possibility is that these features were mentioned as missing from the initial summary because the task of the experiment was to draw the graph from the summary. This observation is reflected by the comments, such as:

"I feel like I had too many "ups and downs". Maybe try to give a better estimate of how many peaks there are"

when talking about the volatility of a trend that he/she was trying to draw. However, the generated summary's intention was to provide a blind user with the information that the trend was not steady, but instead had fluctuations.

The validation of the choice of a different organization structure (not necessarily describing the graph from left to right) could be achieved through an experiment with visually impaired users, discussed in detail in Chapter 8. That experiment assessed the influence such decisions had on their understanding of the high-level knowledge about the graphic conveyed by the summary.

5.8 Thought Experiment

The lack of a baseline for the system kept us from comparing the results from human judgements on the baseline against the system's output. One possibility could be to evaluate how well PageRank and the stopping criteria performed by comparing the results of it with a random number of top N highest initial importance scores of

nodes (as defined by the experiment described in Chapter 4), before running PageRank.

5.9 Summary

This chapter discussed the content determination phase of the SIGHT system. It showed how a proposition's chance of being selected for inclusion in summaries is affected by its measurements and by the previously selected propositions. The chapter describes how a graph-based algorithm is adapted in order to address the need for reflecting different initial weights of propositions, as well as their semantic relationships with propositions that were already selected to be part of the summary.

The chapter finally draws conclusions about the approach chosen to select content in the SIGHT system by analyzing results of the evaluation experiment performed with human judges. The experiment was performed in two phases and was crucial for both improving the content determination phase and also for foreseeing some needed actions in the following phases of the NLG pipeline.

Moreover, it allows the possibility of developing general purpose extraction of content after the PageRank graph is properly adjusted to a given knowledge domain. In our case, nodes in the graph are propositions representing the weighted features discussed in the previous chapter, their relationships follow a logical set of hierarchical and complementary rules, and the initial weight (initial importance score in the context of PageRank) of each node was defined based on how frequent the graphical features were mentioned in the experiment in (Greenbacker et al., 2011).

. The next chapter describes the text organization phase of the NLG pipeline in the SIGHT system.

Chapter 6

TEXT ORGANIZATION PHASE

This chapter describes the text organization phase of the generation module in the SIGHT system. It starts by listing some of the existing research and methodologies used to organize text in NLG systems. It further explains how the intended message and other candidate messages, as well as salient visual features, affect the way the summary of a line graph is organized.

The overarching organization principle divides the selected propositions into three groups. The first is the set of introductory, overall information about the graph. The second group details the trends of the graphic and its characteristics. The last group provides computational information over the whole graph.

The organization of propositions in the second group is affected by the importance of a trend and its selected characteristics provided by the graph-based content determination algorithm. Since the importance of a proposition might affect the organization of the summary, the importance value from the PageRank algorithm described in the previous chapter is used as input for the organization phase.

6.1 Related Work on Text Organization for NLG Systems

What information to communicate, when to say what, and which words and syntactic structures best express the desired intent, are the three classes of decisions that constitute the full range of the language generation problem (K. McKeown, 1992). In order to produce comprehensible text, one needs to decide which ordering of

sentences will be most effective in achieving the goal of making the discourse coherent. Organization choices are heavily influenced by the content that is available, by the main task of the system, as well as by users' expectations when the purpose of the system is to answer questions.

A schema based approach to discourse structuring was proposed by (K. R. McKeown, 1985); it identified certain discourse patterns that facilitate different discourse goals. These schemata are used to organize texts that are defined in terms of rhetorical predicates. This approach allows the same content to be organized differently depending upon the discourse goal. Another technique that has been used is top-down planning. This technique organizes the text as a tree-like structure in which the leaf nodes, which are the informational pieces that can have their communicative role identified by the hearer, are connected by inner nodes representing either the relations that identify rhetorical structures (W. Mann & S. Thompson, 1987) or the speaker intention. (Hovy, 1988), (Zhou & Feiner, 1997) and (Moore & Paris, 1993) are some examples of applications of top-down planning for text structuring.

6.2 Organizing the Selected Content of Line Graphs in SIGHT

Because content determination is done prior to the text organization in this work and produces a set of propositions which must be included in the text, a bottom-up methodology to organize the set of selected propositions is most appropriate. Rhetorical predicates are inappropriate since the set of propositions does not comprise a vast domain of discourse goals on which schemata based planning would be favored. Instead, these propositions would be considered "information" (W. C. Mann & S. A. Thompson, 1987). In the domain of line graph summarization, although a different

subset of propositions pertaining to different pieces of the main entity of the graph (entities being the whole graph and its individual trends) might be selected for each graphic, all the information that is available belongs either to the whole graphic (or main entity) or to its sub pieces (the trends). (Demir, 2010) uses three classes to which propositions are assigned in order to organize summaries and follow up responses for bar charts. These classes are **message_related**, **specific** and **computational** and the classes appear in this order in the summaries. The author describes the classes as (Demir, 2010):

"*The message related class contains propositions that convey the intended message of the graphic. The specific class contains the propositions that focus on specific pieces of information in the graphic, such as the proposition conveying the period with an exceptional drop in a graphic with an increasing trend or the proposition conveying the label and value of a salient bar. On the other hand, propositions in the computational class capture computations or abstractions over the whole graphic, such as the proposition conveying the rate of increase in a graphic with an increasing trend or the proposition conveying the overall percentage change in the trend.*"

This work uses the same overall structure for organizing the propositions selected for the summary of line graphs:

- Our first set of propositions is motivated by Demir's **message related**, but contains additional information which introduces the graph. The introduction of the graph consists of its type (line graph), the description of the measurement axis, and the assessed visual features associated with the whole graphic.
- For the **specific** class, (or the second set of propositions in the overall organization architecture) which describes the smaller pieces that comprise the whole line graph, we add detailed information about the graph's trends.

- For the **computation** class, the set of propositions is also about the overall graphic; these propositions provide information about the graph as a whole but are concerned with values, specific dates and deltas. These are, for example, the overall behavior the graphic presents (if the graphic shows an overall increase/decrease), the overall percentage or absolute change, and/or the overall time span of the graphic.

The second group of propositions, the one which describes the details of the line graphs, may consist of propositions/features associated with a number of trends, each having many attributes that should be mentioned. The question that arises is: how to order this information? Taking into consideration that the answer to this question might vary based on the information conveyed, it was decided that the system should present this information in two major structures:

1. In cases where trends in the graph differ in importance, describe the trends in order of importance (with the most important trend presented first)

2. Describe the trends in the order that they appear in the graphic (which applies when all the trends have similar assessed importance);

In case 1, one (or more) trends in the graphic is/are considered more important than the others. In such cases, the claim is that discussing this/these trend(s) first provides the user with the more important information first, followed by complementary information. Alternatively, if all trends are relatively equal in their importance rating, they will be described in left-to-right order. Section 6.3 explains how the importance of a trend is determined.

In both cases, visual features that are selected by PageRank and have a relationship between them (complement or contrast) are organized in such a way that they are conveyed close to each other. The organization rules consider the semantics of the propositions when sending the organized set to the aggregation module.

6.3 Assessing the Importance of a Trend

A trend can stand out in importance based on two different scenarios. The first scenario occurs when there are candidate messages with non-trivial probabilities (besides the intended message). If the trend(s) are part of such a candidate message, they will be described first. Table 6-1 lists the possible scenarios for organization of propositions that highlight a trend.

Table 6-1: Scenarios for organization of propositions highlighting a trend.

	Description	**Consequence**	**Figure with graph example**
Scenario I	A candidate message highlights a trend that is a part of an intended message that has multiple trends.	Order the important part first.	Figure 6-1
Scenario II	A candidate message with multiple trends contains the single-trend intended message.	The candidate message is introduced first; it is indicated that it does not stand alone, but rather is part of another message.	Figure 6-2

The graph in Figure 6-1 is an example in which there is a candidate message (Big Fall) with non-trivial probability that is part of the intended message (Changing Trend). In this case, the falling trend is described first as shown in the summary following the figure.

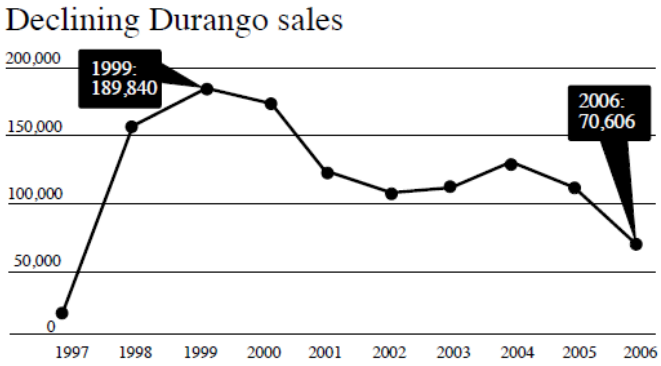


Figure 6-1: Example of Scenario I: a graph with a candidate message (Big Fall) that is part of the intended message (Changing Trend).

(Summary of graph in Figure 6-1)

The image shows a line graph which presents the number of Durango sales. The line graph shows a trend that changes consisting of a rising trend from 1997 to 1999 followed by a falling trend through 2006. ***The falling trend has a starting value of 189840 and has an ending value of 70606. The rising trend is steep.***

Another situation is when the candidate message contains the intended message (the intended message of a big fall is part of a candidate message such as a changing trend that returns, i.e., the falling trend in the big fall is one of the trends in the changing trend message). An example of this is shown in Figure 6-2. This graph has as its intended message the Big Fall, which is part of a candidate message with

non-trivial probability (Changing Trend Return). The summary for this graph is shown below the figure.

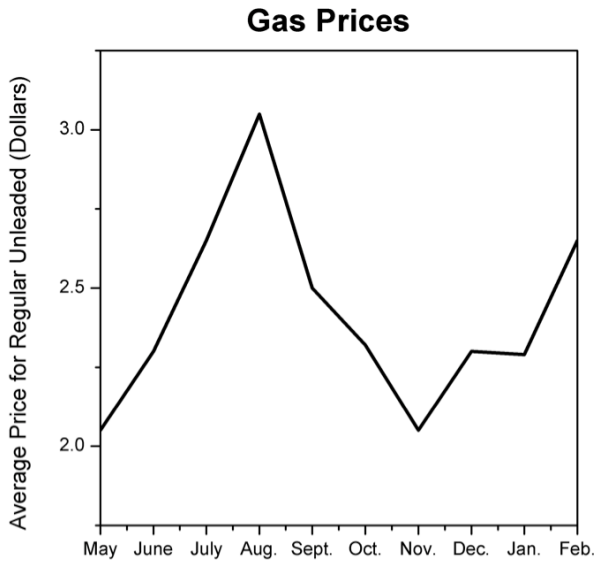


Figure 6-2: Example Scenario II: a graph where the intended message (Big Fall) is part of a candidate message with non-trivial probability (Changing Trend).

(Summary of graph in Figure 6-2)

The image shows a line graph which presents the dollar value of 12-month average for regular unleaded. ***The graph consists of a big fall from 9/4/2005 to 12/4/2005 which belongs to a changing trend that returns. The big fall spans over 90 days. The first segment is a rising trend, which is steep. The third segment is another rising trend.***

The second scenario is dependent on the number and type of propositions selected by the content determination algorithm. In cases where a trend contains more

visual features selected to be included in the summary than any other trends selected for inclusion in the initial summary, the content determination algorithm has implicitly rated that trend as more important than the others. The hypothesis is that that trend would probably be outstanding in the graphic, possibly catching a sighted user's attention before s/he can even read the other trends in the graph, even if they occur before it.

Systems that present time series data such as TREND (Boyd, 1998b), usually organize information based on time. It is crucial to notice that, for systems whose goal is to provide an overall view of data over time, organizing the text according to the order in which they occur might be important in order to consider the evolution of the entity being monitored and draw conclusions. In the case of the SIGHT system, however, summaries should be able to deliver the most important pieces of the graphic first. Our contention is that such pieces are perceived first (if not alone) when the reader is glancing at the graph. These important pieces might occur in the middle of the graphic, for example, and be preceded and followed by some complementary information.

The experiment performed with visually impaired users described in detail in Chapter 8 validates the use of this type of organization by showing that focusing on the most important features, even if not in left-to-right order, was effective for users to be able to answer important questions about the graphic.

6.4 Summary

This chapter presented the organization phase of the SIGHT system. It discussed some of the existing work on text organization for NLG systems and detailed the steps taken for the organization of the propositions selected for a line

graph in SIGHT. The chapter explained the different ways the SIGHT system can organize the selected content and how the choices for one approach over another are made.

The next chapter presents the micro planning phase of the system. Once propositions are selected and ordered, the micro planning phase decides how they should be aggregated and which lexical items should be chosen to describe concepts in the context of line graphs.

Chapter 7

MICRO PLANNING PHASE

This chapter describes the main contribution with regards to the Natural Language Generation aspect of this work. The micro planning phase, which comprises the aggregation of propositions into sentences and the choice of lexical items to describe concepts, is a complex and crucial step when generating text. For this work, the micro planning phase is guided by the generation of text at different reading levels. Studies performed throughout this work showed that tailoring the generated text to a complexity with which the reader is familiar increases understanding and comfortability when reading it.

The chapter starts by describing the motivation behind the choices made for the micro planning phase within SIGHT. It lists the set of propositions and all of the different ways they can be realized. It then discusses related work in both the aggregation and text simplification areas.

In determining the amount of aggregation applied to a summary, the text complexity of the text in which the graphic occurs is taken into account. Grammatical aspects that affect the text complexity are learned in order to guide the aggregation phase so that the realized text readability level matches the readability level of the surrounding text. A graph search technique is employed in order to allow good performance when employing the aggregation step, since so many possible ways of realizing the sentences exist.

This chapter also presents the lexical choice step of the micro planning phase. Once again, the desired reading level guides the choice of the lexical items. To collect a set of relevant terms for describing line graphs, a synonym expansion phase and a word sense disambiguation phase are applied. Ensuring that the synonyms expanded were both relevant to the domain in question (description of line graphs) and appropriate to the different reading levels imposed interesting challenges. These challenges are presented in this chapter. The chapter ends by presenting an example of summaries of a line graph generated at the different reading levels. The evaluation of text complexity is presented in the following chapter.

7.1 Why do NLG Systems Need a Micro Planning Phase?

Deciding on the complexity of a generated text in NLG systems is a contentious task. Some research efforts propose the generation of simple text for low-skilled readers (Williams & Reiter, 2005a); some choose what they anticipate to be a "good measure" of complexity by balancing sentence length and number of sentences (using scales such as the D-level sentence complexity) for the text (Demir et al., 2008); others target high-skilled readers. In this work, we employ an approach that aims to leverage the experience of the reader when reading generated text by matching the syntactic complexity of the generated text to the reading level of the surrounding text. We propose an approach for sentence aggregation and lexical choice that allows generated summaries of line graphs in multimodal articles to match the reading level of the text of the article in which the graphs appear.

As presented in (Moraes, McCoy, & Carberry, 2014b), the main motivation for choosing a scalable approach for this phase is based on the need identified in current generation systems for a generic approach that allows a system to adapt the generated

text complexity to a given user context. This work was motivated by the desire to enable NLG systems to adapt their generated text to different levels of text complexity upon identification that readers will have better comprehension when reading text that is generated at their reading levels. Additionally, if the generated text for the graph matches its surrounding text with regards to readability level, we ensure accordance between their levels, therefore avoiding abrupt changes. Our method for adaptation takes into consideration grammatical features as well as lexical choice[6]. This chapter details the steps taken in order to achieve this capability in the context of line graph summary generation within the SIGHT system.

Initially, in the context of this work, the micro planning phase, which comprises lexicalization and aggregation, was designed and implemented using a rule-based approach in which different text plans were designed for creating summaries for the different grade levels. The text plans were defined by assessing the upper bound of the reading level if all the propositions about a graph were selected. By having all of the possible propositions available for creating the text plans for the different sizes of graphs (1, 2, 3 or more trends), the system would not generate text that would be more complex than desired in cases where the content determination algorithm selected more propositions. In contrast, the current version of this work has been improved to tackle this problem by proposing an approach which combines a learning phase, a concept expansion and word sense disambiguation phase, and a graph search

---

[6] It is important to mention that the aspects considered for guiding the generation of text at different complexity levels in this work are related to syntactic, grammatical and lexical ones. Discourse strategies and other linguistics features are not taken into account.

phase in order to be able to generate text at different reading levels. Even though some pieces of the approach are preprocessed offline for lexicon building, the algorithm used to aggregate propositions in order to select appropriate lexical items to reach a desirable text complexity happens during runtime.

The major identified steps needed to achieve the generation of text at different target reading levels encompass: 1) understanding what makes a text passage complex; 2) mapping measurements of text complexity to specific actions when aggregating propositions into sentences; 3) choosing the set of appropriate words to be used when generating a text passage. The following sections will cover some prior work on sentence aggregation and text simplification as well as all the details on the implementation of the steps mentioned above.

7.2 Related Work on Aggregation of Multiple Propositions into Single Sentences

Interesting work has been pursued in the area of aggregation of sentences. The approach proposed by (Wilkinson, 1995) divides the aggregation process into two major steps: semantic grouping and sentence structuring. Although they are interdependent, both are needed in order to achieve aggregation in a text. Initiatives on automatic aggregation (or only semantic grouping) of text using learning techniques also exist. (Barzilay, 2006; Barzilay & Lapata, 2006) uses a database and its attributes to formalize semantic grouping as a set partitioning problem. It automatically learns grouping constraints by using an aligned parallel corpus of sentences and their underlying semantic representation. (Bayyarapu, 2011) presents an algorithm for context sensitive aggregation that learns aggregation rules that take into consideration the context in which concepts are described and related to each

other. It uses a parallel corpus of multi-sentential text and their underlying semantic representations in order to learn. It models the problem of semantic grouping as a hypergraph partitioning problem that uses the probabilities obtained from a context-dependent discriminative model. (Walker, Rambow, & Rogati, 2001) propose SPoT, a trainable sentence planner that generates a large set of possible organizations for a sentence. It trains a ranker that chooses which of the sentence's organizations are preferred based on a corpus of dialogs and feedback provided by human judges.

Although these learning methodologies are innovative, they assume that there is one best way to aggregate the text (based on human judgments). However, the graph summaries that the SIGHT system aims to generate occur in articles and the complexity of the articles' texts vary considerably. Rather than a single "perfect" aggregation level, this work contends that the text in the summary should match as much as possible the reading level of the text of the article in which the graphic appears.

In the version of SIGHT for simple bar charts (Demir, 2010), the aggregation of sentences within each semantic category is done by considering all possible ways the sentences can be aggregated. Her mechanism treats each proposition as a single node tree, which can be realized as a sentence and attempts to form more complex trees by combining trees in such a way so that the more complex tree (containing multiple propositions) can still be realized as a single sentence. In order to decide which tree is the best one to realize, Demir's work tries to balance sentence complexity and number of sentences. It takes into consideration center-embedded and right-branched relative clauses and their different complexity levels. Demir uses the revised D-level sentence complexity scale (Covington et al., 2006) in order to measure

the syntactic complexity of a sentence according to its syntactic structure and make a decision about the best structure to use.

7.3 Related Work on Text Simplification and Readability Assessment

Research on generating text concerned with low-skilled users has been conducted by (Williams & Reiter, 2004, 2005a, 2005b, 2008; Williams, Reiter, & Osman, 2003). As stated by (Williams & Reiter, 2005b), most NLG systems generate text for readers with good reading ability. Thus, they developed a system called SkillSum which adapts its output for readers with poor literacy after assessing their reading and numeracy skills. Their results show that, for these target readers, the micro planning choices made by SkillSum enhanced readability. The work does not consider higher skilled readers.

(Siddharthan, 2003) proposes a regeneration phase for syntactic text simplification in order to preserve discourse structure, with the objective of making the text easier to read for some specific target reader groups or simpler to process by a computer program. (Carroll et al., 1999) presents a text simplification methodology to help language-impaired users. (Rello & Baeza-Yates, 2012) investigates dyslexic errors on the Web and (Rello & Baeza-Yates, 2014; Rello, Baeza-Yates, Bott, & Saggion, 2013; Saggion et al., 2015) proposes a system that uses lexical simplification to enhance readability and understandability of text for people with dyslexia. They help users to understand the text by offering as options the replacement of more complicated lexical items by simpler vocabulary. They performed experiments with people with no visual impairments and with people with dyslexia and other visual impairments and concluded that the system improved readability for the users with dyslexia and improved comprehensibility for users with no visual impairments.

Effort has also been made on evaluating text simplification systems. (Temnikova & Maneva, 2013) presents an evaluation metric that aims to allow comparison across different text simplification systems by creating C-Score, a common evaluation measure. (Stajner, Mitkov, & Saggion, 2014) proposes some automatic measures that aim to evaluate the grammaticality and meaning preservation of the output text of text simplification systems to replace human evaluation.

Although text simplification is crucial when generating text for low-skilled readers and users with language disabilities, experiments performed with college students (described in detail in Chapter 8) showed that the simplest text was rather unpleasant to read for the majority of them. Just as high-level texts are difficult for a low level reader, over simplified texts are disconcerting to a high-level reader. Therefore, this work proposes a technique that focuses on adjusting the generated text to the reading level of the surrounding text. Thus, the new version of the SIGHT system, the product of this work, aims to satisfy both high-level and low-level readers.

Recently, Artificial Intelligence has been applied to systems that aim to assess and predict the reading level of texts. Language models and Natural Language Processing have been used for predicting the grade level of documents. (Si & Callan, 2001) and (Collins-Thompson & Callan, 2004) predict grade levels of documents by training unigram language models. In addition to language models, (Heilman, Collins-Thompson, Callan, & Eskenazi, 2007; Heilman, Collins-Thompson, & Eskenazi, 2008) and (Schwarm & Ostendorf, 2005) use syntactic features to estimate the text's grade level. (Pitler & Nenkova, 2008) additionally looks into discourse features in order to assess text quality for educated adult audiences using texts from the Wall Street Journal as the corpus. (Kate et al., 2010) presents a system developed to

classify the text readability based on syntactic, lexical and language modeling features. The system performed better than naïve judges when classifying documents based on their readability when compared to annotations provided by language experts. Although these efforts also look into classification of texts based on their readability levels, the work in this thesis uses the classification as a means of assessing the values associated with the features in order to use that as the input to an aggregation module in an NLG system.

(Tanaka-Ishii, Tezuka, & Terada, 2010) presents comparators implemented using Support Vector Machines that are used to, given a set of texts, sort the documents by their readability level. An analysis of the usefulness of applying learning algorithms and sophisticated linguistic features (that go beyond the "classic" features used by more established readability measurements) is presented in (François & Miltsakaki, 2012). (Kanungo & Orr, 2009) uses simple surface level features, like the number of characters and syllables per word, capitalization, punctuation, ellipses etc., to train a regression model to predict readability values in the task of predicting readability of web summary snippets produced by search engines.

7.4 What is There to Realize?

Every NLG system is designed to translate concepts into natural language. They start with the set of concepts or ideas that need to be realized (part of the content determination step) and then organize and plan the structure and surface realization of these concepts or ideas.

In this work, the concepts are related to describing information graphics for visually impaired users. The set of concepts, called propositions, represent semantics of this domain. The micro planning phase of the system starts from a set of selected

propositions, which are ordered by the organization phase, and applies aggregation and lexical choice techniques in order to produce the final text. Table 7-1 shows the propositions which are the set of all possible proposition types that can be chosen by the content determination module. Some of them are presented with a simple sentence translation based on the graph example shown in Figure 5-4 (page 71) for illustration:

Table 7-1: List of all propositions that can talk about a graph.

Proposition Type	Sentence Translation
graph_type	The image shows a <graph_type> - line graph. - "The image shows a line graph".
entity_description	The graph presents <entity_description> – "The line graph presents the *value of Dow*". This is a special case where the entity description is the Measurement Axis Descriptor (MAD) identified and produced by the work presented in (Demir, 2010).
graph_volatility	The graph is (<degree>) volatile/smooth. / The graph shows (<degree>) volatility. - In this case, the degree is defined by the metric calculation result. - "The graph is highly volatile." / "The graph shows much volatility."
graph_overall_behaviour	The graph shows an overall (<graph_overall_behaviour>). - "The graph shows an overall increase."

graph_absolute_change	The graph changed by (<graph_absolute_change>) (<unit>).
graph_rate_change	The graph (<graph_overall_behaviour>) (<graph_rate_change>). - "The graph increased by x percent."
graph_overall_period_years	The graph spans over (<graph_overall_period_years>).
graph_overall_period_months	The graph spans over (<graph_overall_period_months>).
graph_overall_period_days	The graph spans over (<graph_overall_period_days>).
graph_initial_date	The graph starts in/on (<graph_initial_date>).
graph_end_date	The graph ends in/on (<graph_end_date>).
graph_initial_value	The graph starts at (<graph_initial_value>).
graph_end_value	The graph ends at (<graph_end_value>).
maximum_point_value	The graph has a maximum value of <maximum_value>.
minimum_point_value	The graph has a minimum value of <minimum_value>.
maximum_point_date	The maximum value occurs at <maximum_value_date>.
minimum_point_date	The minimum value occurs at <minimum_value_date>.

composed_trend	The graph shows a <composed_trend>. – This describes the intended message of the graph - "The line graph shows a trend that changes. "
trend_description	The <segment_position> is a <trend_description>. – "The first segment is a big fall. " Options of trend description are: rising, falling, stable, big fall, big jump, point correlation and non-sustained.
trend_volatility	The <trend_description> is (<degree>) volatile / smooth. – Where the degree is defined by the metric calculation result.
trend_steepness	The <trend_description> is (<degree>) steep / flat. – Where the degree is defined by the metric calculation result.
trend_initial_date	The <trend_description> starts in/on <trend_initial_date>.
trend_end_date	The <trend_description> ends in/on <trend_end_date>.
trend_initial_value	The <trend_description> has an initial value of <trend_initial_value>.
trend_end_value	The <trend_description> has an end value of <trend_end_value>.
trend_absolute_change	The <trend_description> has a total increase/decrease of (<trend_absolute_change>).

trend_rate_change	The <trend_description> increased/decreased by (<trend_rate_change>).
trend_overall_period_years	The <trend_description> spans over (<trend_overall_period_years>).
trend_overall_period_months	The <trend_description> spans over (<trend_overall_period_months>).
trend_overall_period_days	The <trend_description> spans over (<trend_overall_period_days>).

As described in Chapter 5, a subset of these are selected for the initial high-level summaries of line graphs, according to their importance.

The goal of the micro planning phase is to realize the set of selected propositions as sentences. However, there are many ways these propositions can be realized. They can each originate a sentence, some of them can be realized as an adjective attached to a noun phrase, as a noun phrase added to a conjunction with a preexisting noun phrase, or as a subordinating conjunction. The last three realization options require what we call aggregation of propositions, where multiple propositions are composed to form a complete sentence.

The proposition **graph_type**, for example, can originate:

- A sentence: “There is a line graph.”.
- An adjective (or compound noun): “…line graph…” – where “graph” is the head noun.
- A relative clause: “…which is lined…”. – where the head noun is “graph”.

The other propositions can also have their realizations made in different ways (based on grammatical restrictions of how each concept can be described). A hard decision, therefore, is to choose which realization to apply to each proposition. Another important issue that needs to be addressed is lexical choice. What words should be used in the example above? Should we use graph, chart, or diagram? Deciding how to realize a set of propositions requires a complex and hard set of decisions. How much should be aggregated and how, and which lexical items should be used to describe concepts are the key questions this chapter addresses. The approach used for the system in the context of this work is to use the reading level of the article in which the graphic appears to guide such decisions. The idea is to generate summaries in a given reading level so that their reading levels match. How the system employs this approach is the main focus of this chapter.

7.5 Planning the Realization of Propositions

Thus far, it has been shown how propositions can be realized as full sentences. However, we want to be able to aggregate sentences together and we want all of the possible combinations. Now we are going to formulate the problem of generating all of the possible realizations using graph search. By having all of the possible realizations, we will be able to choose the one at the desired reading level. For this to happen, however, we need to learn about text complexity and be able to augment the algorithm by adding that information to a heuristic, so it can inform us how close we are to the desired output. Finally, we want to do this efficiently, so we choose and apply a graph search algorithm that, by using the heuristic, can lead us to the desired output faster.

7.5.1 Realizing each proposition

In Section 7.4 the set of propositions was presented, along with some examples of their realizations as active voice sentences. The possible realizations being considered in the context of this work are:

1) Realization as a sentence in active voice
2) Realization as a sentence in passive voice
3) Realization as a relative clause
4) Realization as an adjective
5) Realization as a conjunction

Not all of the possible realizations can be applied to all of the proposition types and some propositions can have more than one realization under one of these options. One example of the former is the proposition **composed_trend**, which conveys the intended message. This proposition cannot be realized as an adjective as its core description is a phrase on itself (e.g., The line graph shows **a trend that changes**). For this proposition, the possibilities are:

1) Realization as a sentence in active voice: The line graph shows **a trend that changes.**
2) Realization as a sentence in passive voice: **A trend that changes** is shown by the line graph.
3) Realization as a relative clause: … graph, which shows **a trend that changes**, …
4) Realization as a conjunction: …graph, which presents the value of Dow **and shows a trend that changes**, …

The example above also applies to propositions such as **graph_overall_behavior**, **graph_initial_value**, and many others.

Other propositions can be realized as all four alternatives. In fact, they have different ways in which they can be realized within each alternative and these are determined by the lexical item being used. One example of such a proposition is **graph_volatility**. Since this concept can be described by adjectives (volatile, jagged, variable) as well as by nouns (volatility, fluctuation, variability, jaggedness), there are multiple options for realizing this proposition for each alternative[7]:

1) Realization as a sentence in active voice: The line graph is **volatile**. / The line graph shows **fluctuation**.
2) Realization as a sentence in passive voice: **Volatility** is shown by the line graph. (Even though the construction "Jagged is the line graph" is grammatically correct, we decided not to allow such a realization because it is unusual. The example in Figure 7-1 shows the comparison between the occurrence of the sentences "The girl is beautiful" and "Beautiful is the girl" in the Google Books Ngram corpus (Michel et al., 2011), showing that contemporary language has dropped such use, making it unlikely to appear in popular media available online. So the idea is that the system should also take this into consideration when generating summaries of the graphics).
3) Realization as a relative clause: … graph, which is **variable**, … / … graph, which shows **jaggedness**, … - where **graph** is the head noun.
4) Realization as an adjective: …**volatile** graph… - where graph is the head noun.

---

[7] Appendix A shows the formalization of all the different proposition realization templates.

5) Realization as a conjunction: …graph, which presents the value of Dow **and shows much fluctuation**, …

The propositions **trend_volatility** and **trend_steepness** have a similar behavior to the proposition **graph_volatility**. The concept steepness can also be described by adjectives (steep, abrupt) as well as by nouns (steepness, abruptness).

The propositions which share the same set of possible alternatives and the same root predicate (show, present, have) are combined and use the same proposition realization template, where just the concepts and values are instantiated for each individual proposition.

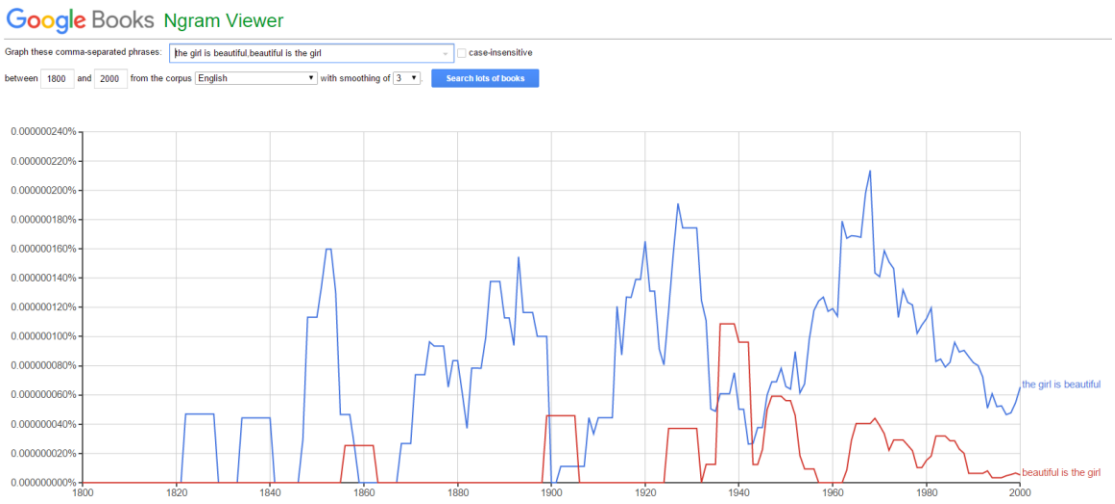


Figure 7-1: Snapshot from Google Books Ngram Viewer (books.google.com/ngrams) comparing the usage of the sentences "The girl is beautiful" versus "Beautiful is the girl".

7.5.2 The graph search problem for realizing propositions

For generating graph summaries at a desired reading level, we are formulating the problem as a graph-search through the space of possible realizations. The following describes the search space: its states, actions, transition model and goal test.

States: A state consists of two parts: a list of unrealized propositions and the realizations performed so far (which can consist of full sentences or sentence fragments).

Initial state: The initial state contains the set of all propositions unrealized.

Actions: The actions in a given state take the next unrealized proposition and realize it (generating a new state for each realization the proposition allows). The possibilities are: *realize_as_active_sentence*, *realize_as_passive_sentence*, *realize_as_adjective*, *realize_as_relative_clause* and *realize_as_conjunction*. Each proposition contains a set of its allowed actions. Figure 7-2 shows a piece of the graph search in which the proposition **graph_volatility** (for the graph example presented in Figure 5-4 page 71) is the next to be realized[8]. It illustrates the states that result from a node containing **graph_volatility** as the next proposition to be realized is chosen from the open list and expanded. If the needed head noun is not present in any of the realizations, then some of the actions (adjective, relative clause and conjunction) will be realized as segments and will wait until such a head noun is generated to be added to a full sentence. If the required head noun is already realized in a full sentence, the fragment is then attached to the existing realization. In the example presented in Figure 7-2 the head noun **graph_type** is present in a full sentence, so the fragments could be attached. The fragment for the relative clause was left out for illustrative

[8] The proposition order is provided by the organization module.

purposes, but it could be added as a relative clause or as a conjunction (generating two different successors).

Goal test: It checks if all the propositions have been realized and if all of them are aggregated into full sentences.

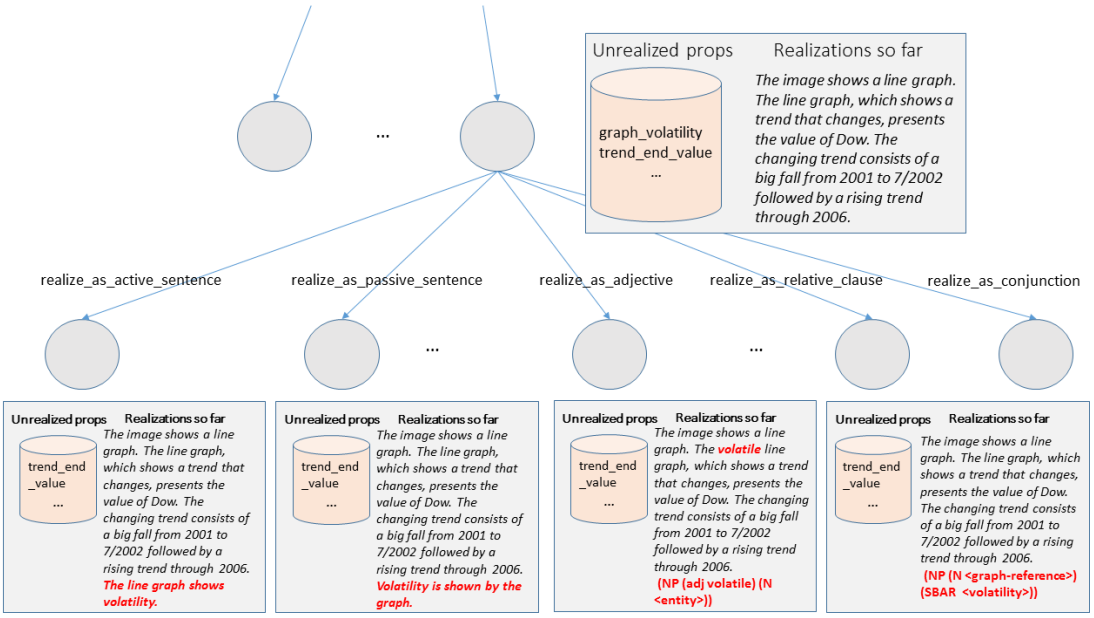


Figure 7-2: Snapshot from the Best First Search algorithm at a point where the proposition **graph_volatility** is being expanded.

The partial summary realized so far and the set of unrealized propositions is used to calculate h(n). For each unrealized proposition, one new node is added to the open list for each possible realization of that proposition. If a proposition can be realized as a single sentence in active voice, a single sentence in passive voice, an adjective and a relative clause, four new nodes will be added to the open list, one for

each possible realization of such proposition. The same applies to the whole set of unrealized propositions.

The graph search algorithm chosen was the Best First Search algorithm (BFS) (Russell & Norvig, 2003). The choice for this algorithm was motivated by the size of the search space and the need for a complete solution. Algorithm Best First Search uses the formula ***f(n) = h(n)***, at each expanded and visited node in the graph – say *n*. For that node, the estimated cost from *n* to a goal node is defined by a heuristic.

In this work we develop a special heuristic, composed of two main factors. The first factor is the level of the node in the tree: favor nodes deeper in the tree (i.e., closer to being fully realized). The second factor considers how likely the eventual realization is to be within the range of features for the desired grade level. This is estimated by taking into account both the realization so far and the estimation of how the features are likely to change given the proposition that have not yet been realized. The next section provides details on the heuristic function.

The next section describes the steps needed to build the graph search algorithm, especially the path taken to learn the measurements that will be mapped to the heuristic.

7.6 Building the Graph Search Algorithm

7.6.1 Finding the heuristic to estimate text complexity

In order to define a heuristic that will allow the implementation of the Best First Search algorithm, one needs to understand which aspects of text contribute to its complexity, identify a subset of these that can be used when generating text, and map its values to functions to construct the heuristic.

7.6.1.1 Understanding what makes text complex

Various reading assessment measures exist today. Given an excerpt of text, these measures usually take into consideration features such as word and sentence lengths, and some syntactic structures in order to assess the grade level the reader should be at in order to easily understand the text. Described next are some of the most commonly used reading level measurement techniques. Here we describe several simple measurement techniques that can be automatically applied to text through freely available software packages.

7.6.1.1.1 Automated Readability Index

The Automated Readability Index (Smith & Senter, 1967) relies on a factor of characters per word (a character is a letter, a number, or a punctuation mark) to assess word length (which is one of the major factors for assessing grade level). Some other measures use syllables per word for their computation of grade level and still others use complex word indices. Although opinion varies about its accuracy as compared to the syllables/word and complex words indices, characters/word is often faster to calculate, as the number of characters is more readily and accurately counted by computer programs than syllables (Smith & Senter, 1967). In fact, this index was designed for real-time monitoring of readability on electric typewriters. The grade level is calculated as:

$$4.71\frac{characters}{words} + 0.5\frac{words}{sentences} - 21.43 \tag{3}$$

7.6.1.1.2 Flesch-Kincaid

The Flesch-Kincaid grade level readability formula improves upon the Flesch Reading Ease Readability Formula (Kincaid, Fishburne, Rogers, & Chissom, 1975). Originally formulated for US Navy purposes, this formula is best suited to the field of education. The two measures (FleschIndex and Kincaid) use word and sentence length with different weighting factors. FleschIndex is a test of reading ease with higher scores indicating text that is easier to read. Kincaid is a grade score that is negatively correlated to FleschIndex and provides a grade level for the text. Formula 4 represents the measure:

$$\alpha * \frac{words}{sentences} + \beta * \frac{syllabes}{words} - \gamma \quad (4)$$

where the weights assume the values −1.01, −84.6 and 206.83 respectively for FleschIndex and the values 0.39, 11.8 and −15.59 respectively for Kincaid.

7.6.1.1.3 Coleman-Liau Index

The Coleman-Liau index was developed with the goal of making the process of assessing the reading level of a text faster (by not depending on the assessment of the number of syllables), since keypunching the text into the computer in order to be able to access syllables was generally more expensive than obtaining a reading ease score by hand counting (Coleman & Liau, 1975). It proposes the use of an optical scanner to count all words occurring between two periods. The authors claim that it would be equally simple for the same device to count word length by requiring that word length be measured in letters, not syllables. The index is represented by a cloze score (instead of a grade level) which, according to the authors, might not be as easily

readable by users. The cloze score can be converted to a grade level using a second formula.

$$0.059\,L - 0.296\,S - 15.8 \tag{5}$$

where, L is the average number of letters per 100 words and S is the average number of sentences per 100 words.

7.6.1.1.4 SMOG (Simple Measure Of Gobbledygook)

The SMOG grading level assessment uses the number of polysyllables (a word consisting of more than three syllables) and the number of sentences in order to calculate the grade level of a given passage. According to (Laughlin, 1969), the SMOG grade yields a 0.985 correlation with a standard error of 1.5159 grades with the grades of readers who had 100% comprehension of test materials. SMOG has been widely used in health messages. According to the author, word length is associated with precise vocabulary in the English language, so a reader must usually expend extra effort in order to identify the full meaning of a long word, simply because it is precise. In the same way, long sentences nearly always have complex grammatical structure, which is a strain on the reader's immediate memory because he has to retain several parts of each sentence before he can combine them into a meaningful whole. The SMOG formula is given below:

$$1.043\sqrt{numOfPollysyllables * \frac{30}{sentences}} + 3.219 \tag{6}$$

The measurements described above generally do not agree on their assessed reading grade level when they analyze a passage. A tool available in the GNU project Style and Diction (Fsf, 2005) provides results for ARI, Flesch-Kincaid, Coleman-Liau,

Fog, Lix and SMOG. As a baseline, we average these measures for comparison purposes when evaluating the text generated at different reading levels by the system.

7.6.1.1.5 Latest efforts on readability measurement

The previous measures are the ones most widely used. They have been available for a couple of decades and are relatively easy to use. However, findings by the Common Core State Standards Initiative (Common Core State Standards Initiative, 2010) have shown that their capability is limited when analyzing text complexity. Motivated by this, the Common Core Standards launched a challenge for the creation of new measurement tools that would additionally consider textual aspects such as grammatical, discourse-related, and genre related when evaluating and assessing the readability of a passage.

Examples of measurement tools that resulted from this effort are *TextEvaluator* (previously named *SourceRater*) (Napolitano, Sheehan, & Mundkowsky, 2015; Sheehan, Kostin, Futagi, & Flor, 2010) and Coh-Metrix (Graesser et al., 2004). TextEvaluator, for example, uses eight component scores, each of which is a linear combination of four to ten fine-grained features. It follows the Common Core Standards methodology of grouping texts into two types: Informational and Literary. The component scores consist of the following groups: sentence complexity, vocabulary difficulty, connections across ideas, and organization.

In order to use the measurements as a guide for generating text, one needs to know exactly how each feature affects readability. Obtaining such a level of precision becomes a challenge through the analysis of the results provided by these latest tools. Since estimating their individual feature measurements is not straightforward, they are kept from being readily available for use as the input for an external Natural Language

Generation system. Additionally, some of the features used by such tools are not relevant for generating graph summaries (one example is the presence of discourse strategies such as persuasion and negation).

Motivated by these constraints on using available readability measurement tools, a learning approach was developed specifically for this work. The goal was to identify the set of features of language and their weights that directly affect text complexity and, at the same time, can be used as the input for a NLG system in a straightforward manner. This learning approach is described next and it has as its sole purpose to provide the NLG system with a set of weighted parameters that will be applied during the aggregation phase.

7.6.1.2 Learning the importance of a specific subset of features which affects text complexity

As mentioned previously, a learning approach was taken in order to learn which set of features and their weights leads text to have varied complexity levels. These features need to be chosen based on both their effect on text complexity and their usability. The choice of features for constructing the model was made based on the work presented by (Vajjala & Meurers, 2012) which uses Second Language Acquisition (SLA) research based features combined with traditional readability features such as word length and sentence length in order to classify text into different grades. Their work results in classifiers that outperform previous approaches on readability classification, reaching higher classification accuracy. However, since this work still needs to map features back to the NLG aggregation phase, the set of features used by SIGHT represents a subset of the features presented by their work.

7.6.1.2.1 Feature engineering and learning algorithm choice

For the learning algorithm a decision tree is used. This algorithm was selected after an analysis of accuracy and ease of assessment of the features and their values. The goal of the learning algorithm was to provide the system with concrete measures of the chosen features that can be used during the aggregation phase. The set of features, motivated by the work presented in (Vajjala & Meurers, 2012), used to train the model were:

- Percentage of sentences starting with a pronoun (*percBegSentPronoun*);
- Percentage of passive sentences (*percPassiveSent*);
- Percentage of conjunctions (*percConjunction*);
- Percentage of pronouns (*percProunoun*);
- Percentage of sentences starting with subordinating conjunction (*percBegSentSubConjunction*);
- Percentage of prepositions (*percPreposition*);
- Percentage of sentences starting with conjunction (*percBegSentConjunction*);
- Percentage of nominalizations (*percNominalization*);
- Percentage of sentences beginning with prepositions (*percBegSentPreposition*);
- Percentage of adjectives (*percAdjective*);
- Percentage of adverbs (*percAdverb*);
- Percentage of relative clauses (*percRelativeClauses*);
- Average noun phrase length (*avgNounPhraseLength*);
- Average verb phrase length (*avgVerbPhraseLength*);

- Average sentence length (*avgSentLengthWord*);

7.6.1.2.2 Corpus of grade level annotated text

Data was obtained from text exemplars classified at different grade bands available in Appendix B of the Common Core State Standards ("Common Core State Standards Initiative," 2014) and various articles written and annotated at different reading levels. Magazine articles collected from the Austin Public Library electronic catalog (Library, 2015) were annotated using the Lexile measure ("Lexile Framework for Reading," 2015). Since the Lexile measure uses a different measurement scale (the output is not in terms of grade levels) for selecting articles for different grade levels, a conversion table (shown in Table 7-2) was used. Classes for the learning algorithm were defined as groups of grade levels. The grades were grouped as 4^{th} and 5^{th} grades, 6^{th} through 8^{th}, 9^{th} and 10^{th}, and 11^{th} and up. One hundred articles, varying in size, were collected for each one of the grade level groups. These articles were in HTML format and they were preprocessed to remove tags and special characters. After preprocessing the files, they were split into smaller passages, of at least 150 words, which is equivalent to the average size of the summaries the system generates. Because the passages needed to have complete sentences in order to obtain more accurate measurement of the features during learning, the splitting step counted words sentence by sentence and, after reaching 150 words, it stopped adding sentences to the current passage. Splitting the articles resulted in 1874 passages, which were used as instances in the learning algorithm.

Since some of the articles would have a Lexile measure that belonged to multiple of our grade level groups, only articles that were annotated with Lexile measures belonging to unique grade level groups were used. For example, when

looking at Table 7-2, for the grade level group 4th and 5th grades, the best Lexile measurement representing these in our scenario would be 700 since measures in the upper 600's can also be classified as 3rd grade and in the lower 800's can also be classified as 6th grade. Only articles annotated with 700 Lexile measurement were chosen for that grade group. Similarly, for 11th grade and up, only articles annotated with a Lexile measurement of above 1150 were chosen. After splitting the articles into similar passage sizes, the values of the features are calculated using the Style & Diction tool (Fsf, 2005) for assessing some of the syntactic features and NLTK (Loper & Bird, 2002) for grammatical features. After all the features were assessed, a tab file (appropriate input file type for use with the Orange toolbox (Demsar et al., 2013)) is generated and ready for training. The next section details the learning process, the choice of the right learning algorithm, and the accuracy for the classification of passages.

7.6.1.2.3 Learning algorithms and classification task

Before choosing decision trees as the learning algorithm to be used for this classification task, other algorithms were analyzed using the data described in the previous section and their results were compared. Random forests, Bayesian networks, Classification (or decision) trees and Neural Networks were applied to the classification task. The Orange toolbox was used for this comparison as it makes available a number of different classification, regression and association algorithms for machine learning and data mining. Using leave-one-out cross validation, the system achieved a classification accuracy of 85.38% and F1 measure of 87.97% using decision trees. The Neural Network outperformed the classification accuracy of the decision tree by 1.39%, but had a smaller F1 measure. The neural network used 20

hidden layers, which would probably complicate reading the feature weights due to the combination functions that happen within the hidden layers, for example. Since the goal is to be able to map the weights of the features to a heuristic in a graph search algorithm, the best option turned out to be the decision tree since it provides rules which allow the values of the features to be captured.

Decision trees provide a set of logic rules, which establish a relationship between the features that contributed to the classification and their corresponding values. This is the input used to guide syntactic and grammatical decisions during the micro planning phase. Learned values of these features allow the micro planning phase to use this information to decide on the structure of the generated text.

Table 7-2: Conversion table for Lexile measurement scale.

Grade Level	Accelerated Reader	DRA Level*	Lexile
K	1-1.2	A-2	Beginning Reader
1	1.2-1.4	2	100
		6	
	1.5-1.7	10	200
		14	
	1.8-2.1	16	300
2	2.2-2.6	18	400
		20	
	2.7-3.2	24	500
3		28	
	3.3-3.9	30	600
		34	
		36	
4		38	
	4.1-4.7	40	700
		42	
5	5.0-5.8	44	800
6	6.0-7.0		900
7-8	7-8.9		1000
9-12	9-12.9		1150

Evaluation Results

	Method	CA	F1	Prec	Recall
1	Classification Tree	0.8538	0.8797	0.9070	0.8540
2	Neural Network	0.8677	0.8464	0.8692	0.8248
3	Naive Bayes	0.8111	0.8104	0.8258	0.7956
4	Random Forest	0.7924	0.8310	0.8027	0.8613

Figure 7-3: Comparison across different learning algorithms.

The paths from the root to the leaves (or classes, in this case) provide logical rules that represent the values of the different features which led to that classification. The logic rules can be read as $path_1$ OR $path_2$ OR … $path_N$ for a given grade level group (grade level groups are the target classes of the leaf nodes). Within a leaf node, however, there is a combination of constraints that are satisfied in the path from the root to the leaf which are the values of features in that path. Some classifications have lower confidence than others. Only nodes with a classification confidence above 70 percent were used to construct the set of logic rules that is used by the system. A set of rules for a 9^{th} – 10^{th} grade level band is shown here as an example of what the decision tree produces:

(avgParagLengthSent <= 10 AND (avgSentLengthWord > 13 AND avgSentLengthWord <= 15) AND percPassiveSent <= 0.4 AND numberRelativeClauses <= 0.6 AND percBegSentPronoun > 0.2 AND percBegSentPronoun <= 0.5)

OR

(avgParagLengthSent <= 9 AND (avgSentLengthWord > 14 AND avgSentLengthWord <= 16) AND percPassiveSent <= 0.1 AND numberRelativeClauses <= 0.8)

We use these rules as a heuristic to help guide the search to a realization that satisfies the target reading level. When using these rules within our heuristic, the function will be estimating the cost based on how well the to-be-realized propositions fall within those ranges in order to be inside the grade level constraints.

7.6.1.3 Mapping the rules to a heuristic function

In calculating the heuristic, two groups of features have their costs estimated differently. The **first group** contains features that do not fluctuate their values as new propositions are realized. One example is the number of relative clauses in a paragraph. As the number of sentences in the paragraph increases, the value of this feature can never go down, it only goes up. The **second group** contains features whose values can fluctuate (either up or down) as new propositions are realized. The average sentence length in words, for example, can go up or down as new propositions are realized since they can become new sentences (making it go down) or be aggregated with existing sentences (making it go up). For this reason, the heuristic calculates the estimated cost that is added to *h(n)* differently for these two groups. The next sections explain how they are calculated.

Another aspect, which is common to both groups and is part of calculating the heuristic, is called "depth measure" and it represents the proximity of the current node to a goal node. Since all of the goal nodes are at the same depth in the tree (if there are 8 propositions to be realized, all the goal nodes will be at depth 8), this measure favors the nodes which are deeper in the search when compared to shallower ones.

This cost is obtained by applying formula (7) and it gets smaller as the node gets closer to a goal node. A depth measure of 0 is added when the number of unrealized propositions is equal to 1. When the number of unrealized propositions is equal to 0, a goal node was reached, so h(n) is no longer needed.

$$depth_{measure} = 1 - \frac{1}{unrealizedPropositions} \quad (7)$$

7.6.1.3.1 Calculating the cost added by feature values that do not fluctuate

Features that are part of this category are: *begSentPronoun, numberArticles, begSentArticle, numberRelativeClauses, numberAdjectives, and numberAdverbs*. To illustrate, consider the following example used to explain the heuristic calculation: suppose that the decision tree learned that, for paragraphs that contain around 150 words, the range of values for the *numberAdjectives* feature is 2 <= *numberAdjectives* <= 5 for a 4th grade level text. The sequence of rules to calculate the cost for this type of feature is:

1. **If the measured value of the feature in what has already been realized is above the upper limit of its range** (if it is equal to 6 for the example above)**, add an infinite cost to the estimation.** Since these feature's values can never go down, this node cannot satisfy the requirements for the grade level and so should be ordered towards the end of the open list.
2. **If the measured feature is within the predefined range** (if it is equal to 3 for the example above), **add to the estimation the probability of increasing the value of the feature based on the unrealized**

propositions. In this case, the probability of increasing the feature is the ratio of possible realizations that increase the feature's value (e.g. a proposition that has a possible realization as an adjective will increase the *numberAdjectives*) over all possible realizations amongst the set of unrealized propositions. With this metric, we want to make sure that the nodes that have a higher probability of staying within the range have a smaller cost added to them. In the example above, if there were **6** unrealized propositions from which **2** could only be realized as active voice sentence and passive voice sentence (**4** possible realizations), **1** could be realized as active voice sentence, passive voice sentence and relative clause (**3** possible realizations), and **3** could be realized as active voice sentence, passive voice sentence, adjective, and relative clause (**12** possible realizations), the number of possible realizations would be 19. Since only 3 could be realized as an adjective, the probability of increasing the value of this feature is 3/19 (~ 0.16). This value would be added to the cost, versus 0.31 (6/19) if there were 6 possible realizations as adjectives in the set of all possible realizations.

3. **If the measured value is less than the lower limit** (if it is equal to 1 for the example above), **multiply the probability of increasing the value of the feature given the unrealized propositions** (as explained above) **by the inverse of the value that the feature can increase by** (feature upper limit – feature value = 2 for the example above)**, then multiply the result by the number of possible realizations that use**

the feature. In this case, the more chances to realize a proposition as an adjective the better since the value is currently lower than desired. For the example provided above, 1.44 would be added to the cost for the first case (3 / 19 * 1 / (5-3) * 3) and 5.58 would be added to the cost for the second case (6 / 19 * 1 / (5-3) * 6). The first case is preferred since the second case could go over the upper limit if all 6 possible realizations as adjectives were indeed realized as adjectives. The formula for adding the estimated cost is:

estimated_cost += probabilityOfIncreasingFeature (1 / featureUpperLimit - featureMeasure) * numberUnrealisedPropsAsFeature;*

7.6.1.3.2 Calculating the estimated cost added by feature values that fluctuate

Other features, such as *averages* and *percentages*, need a different logic in order to estimate the cost. This group encompasses the remaining features. Consider *avgSentLengthWord* as an example. As new propositions are realized, the average number of words per sentence can increase or decrease. Therefore, a different logic is followed in order to take this aspect into account. For this type of feature, some assumptions are made regarding the maximum number a proposition can contribute to increasing the value of the feature. For the feature *avgSentLengthWord*, for example, we assume that the maximum number of words that can be added by the realization of any proposition is 15 (this number was chosen by looking at the longest description in the set of propositions for describing the intended message of line graphs). The rules to calculate the estimated cost for this type of feature are:

1. **If the measured value of the feature is within the predefined range**:

- Calculate the **expected increase in feature value.** For the feature *avgSentLengthWord,* for example, this is represented by the number of words that can be added without new sentences being added. This is calculated by summing the probabilities of increasing the number of words multiplied by the number of unrealized propositions, divided by 1 minus the probability of increasing the number of sentences (which is the probability of NOT realizing propositions as single new sentences).
- Add the **expected increase in feature value** to the **current feature value** acquired from the already realized propositions and divide it by the **current number of sentences** to find the projected feature value.
- If the projected feature value is within the range, add the projected value to the estimated cost. If the result goes beyond the limits of the feature, use the difference to calculate the penalty to be added to the estimation. The penalty is represented by the number it went off by times the inverse of the number of realizations (since the greater the number of unrealized propositions the better in this case, since it gives us more room to get to a good final average within the range).

2. **If the measured value is above the upper limit, calculate the probabilities of increasing the feature by the possible exceeding number**. For the *avgSentLengthWord*, for example, accumulate the probabilities of increasing the feature by the numbers of words that are

greater than the limit (if the limit is 5, it will sum up the probabilities of adding 6, 7, 8, 9, 10…15 words), **multiply it by the probability of increasing the number of sentences** (by looking at how many unrealized propositions can be realized as sentences in active or passive voice), then **multiply the result by the number of unrealized propositions**. This is the penalty that should be added to this node if it is already out of limits.

3. **If the measured value is less than the lower limit, calculate the probabilities of increasing the feature by numbers that are below the lower limit. S**ince the average can also change the measured feature to fall lower than the lower limit, we need to address this case. For that, it accumulates the probabilities of numbers of words to be added that are less than the limit (if the limit is 4, it will add the probabilities of adding 0, 1, 2, and 3 words). After summing the probabilities, **multiply it by the probability of also increasing the number of sentences**. Finally, **multiply the result by the number of unrealized propositions**. This is the penalty that should be added to the cost estimation if the measured value of the feature is less than the lower limit.

The final value for h(n) is the sum of all estimated costs when going through the set of unrealized propositions. The calculated value for h(n) is then stored in the node and the priority queue (implemented with a min heap) used to order nodes from the open list uses this value to insert the new node.

7.7 Lexical Choice for Generating Summaries at Different Grade Levels

Work has been proposed to automate the replacement of lexical items for text simplification systems. Rello (Rello et al., 2013) proposes an approach that lists a set of simpler words for the user to choose from in order to replace words considered hard or difficult. It uses a Spanish thesaurus in order to come up with a list of synonyms and word frequency for determination of simpler synonyms (assuming simpler words are seen more frequently). (Saggion et al., 2015) presents a text simplification system which implements a rule-based approach that aims to address textual simplification operations that could not be addressed by synonym substitution.

Lexical choice is the other important piece of the puzzle when it comes to text complexity and grade level. The system needs choices for lexical items and needs to be able to select them such that the generated text is at the desired grade level.

The first thing that comes to mind when one needs to find variations of concepts from which to pick from is to perform synonym expansion. So the first attempt was to use a thesaurus in order to collect synonyms of the terms used to describe the concepts in the system. By searching Thesaurus.com (Dictionary.com, 2015) one can notice that synonyms are grouped based on *synsets* – similar to the way lexical items are grouped in the WordNet database (Fellbaum, 1998). For the concept **trend**, for example, the thesaurus provides two different *synsets* – one for the concept of *flow, current* and another for the concept of *style, fashion that is in favor*. Thus, one option could be to choose one *synset* from which the synonyms could be used. However, two problems were found with this approach: 1) there were no comprehensive *synsets* which fairly covered the set of appropriate lexical items for all of the concepts that could be used in the context of describing line graphs. The *synsets* would either have a large number of inappropriate synonyms or the appropriate

synonyms would be scattered across different *synsets*; 2) by choosing a *synset* manually, the lexical choice module would be implementing a highly supervised approach, which was not what was desired. The solution was to devise a set of techniques that would allow the system to, for each concept, **create a relevant *synset* for the domain in question by using all of the available synonyms** (given a specific part of speech) **and further apply a word sense disambiguation step based on a pre-defined context for each concept being described**.

After defining its own *synset* for each concept, the system needs to be able to determine which word(s) should be used to describe a concept for each of the different target reading levels. These phases are described in the next sections of this chapter.

7.7.1 Concept expansion phase

The first step in constructing a pool of synonyms for choosing from in order to realize concepts that occur in a proposition is to find a base lexical item to represent the concept and that can be expanded by collecting synonym. The first challenge was to decide which term would be the base lexical item. For the volatility concept, for example, one can start by expanding the term "jagged" or the term "volatile". Either of two terms would describe the idea satisfactorily for line graphs, but how should we decide which lexical items are used for each concept? Since the thesaurus used in this work presents different synonyms for these terms, we wanted to rely as much as possible on the way people would describe this concept. For this reason, the base lexical items for each concept were gathered from the experiment performed by (Greenbacker et al., 2011) in which participants were asked to describe the important aspects they noticed were present in the line graphs. From these passages, the most

common words used to describe concepts such as volatility and steepness were used as the starting point for lexical building.

For expanding these concepts, Thesaurus.com (Dictionary.com, 2015) was used. Thesaurus.com was selected because it has a better coverage with respect to synonyms of nouns, verbs, adjectives and adverbs than WordNet (Fellbaum, 1998) and VerbNet (Kipper, Dang, & Palmer, 2000). Thesaurus.com provides synonyms for concepts in a varied number of senses and parts of speech by grouping synonyms within part_of_speech + *synsets*. As mentioned previously, choosing the most appropriate concept *synsets* for the domain of line graphs did not appear to be the best approach, as the *synsets* were not always comprehensive and precise. In other words, all synsets individually contained some synonyms which were not appropriate and appropriate synonyms were found across multiple *synsets*. Besides, choosing a single best *synset* would not lead to a technique that could perform the synonym expansion without human supervision. For this reason, the decision was therefore to use all *synsets* with a given part of speech and to further filter the resulting set. The algorithm uses all of the synonyms, in all the different *sysets* (for the same part of speech being applied to the concept by the system) present in the Thesaurus.com website as the first step for concept expansion.

This provided the system with an extensive (and noisy) list of synonyms. The set of synonyms was too broad; it included synonyms that would not apply at all to the domain of line graph description. One example is the expansion of the concept **show**. Figure 7-4 through Figure 7-7 extracted from the Thesaurus.com website show the *synsets* for the concept **show** used as a verb.

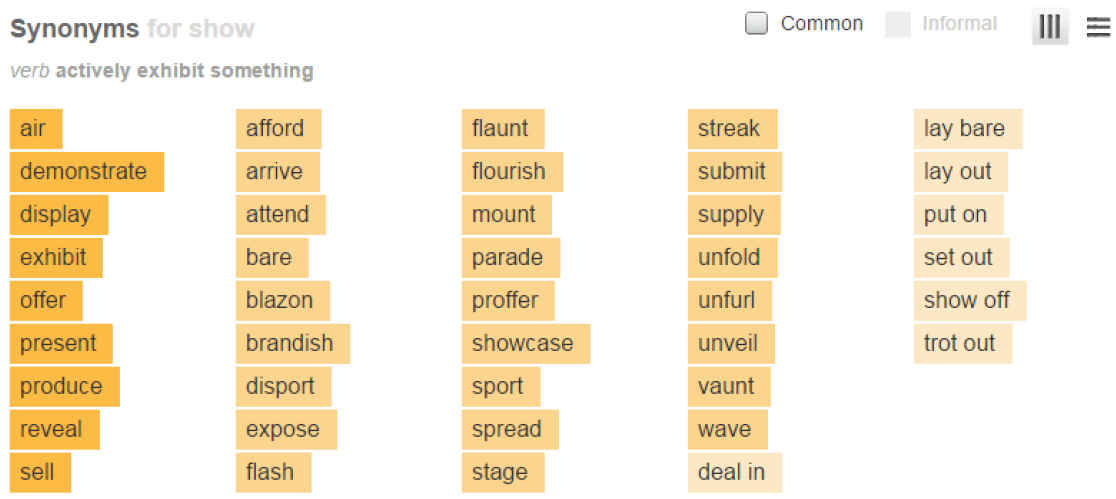


Figure 7-4: Thesaurus' synonyms for concept **show** used as a verb with a sense of "actively exhibit something".

Figure 7-5: Thesaurus' synonyms for concept show used as a verb with a sense of "passively exhibit something".

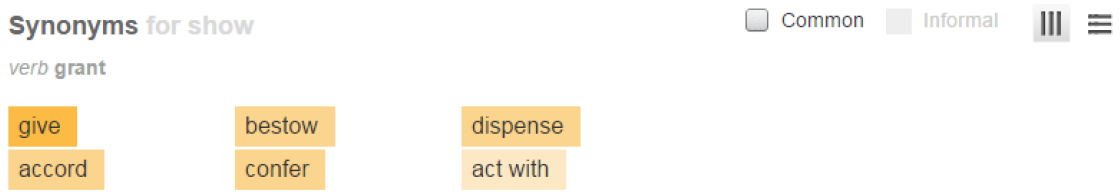


Figure 7-6: Thesaurus' synonyms for concept show used as a verb with a sense of "grant".

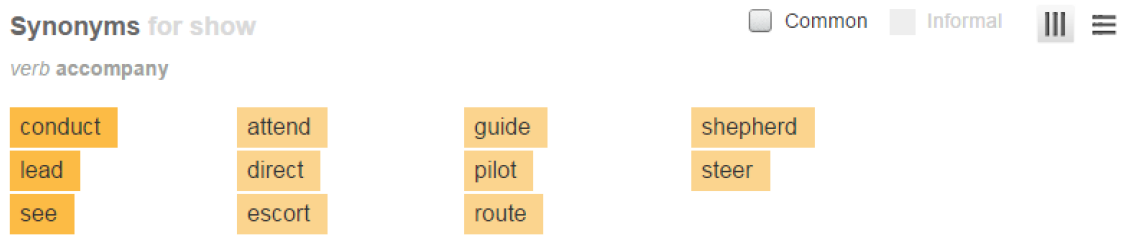


Figure 7-7: Thesaurus' synonyms for concept show used as a verb with a sense of "accompany".

For example, in the sentence “The image shows a line graph”, many synonyms would not sound good (“The image **pilots** a line graph” or “The image **bestows** a line graph”). On the other hand, either of “The image **reveals** a line graph” or “The image **illustrates** a line graph” would be appropriate lexicalizations of the concept **show** in this context.

The system aims to find the subset of synonyms of a concept that would be appropriate in the context of line graphs; therefore, it employs a set of approaches that, combined, allows the creation of a “domain aware *synset”* which is concerned with the context of the domain in question (line graphs, in this case).

7.7.2 Disambiguating the set of synonyms for the line graph domain

7.7.2.1 Language modeling: Using 5-grams from the Google Books corpus

This work applies language modeling to filter the synonyms of a word to that subset that is applicable in the context of line graphs. The intuition is that we want to keep only those synonyms that the language model indicates appear in a context containing key words indicative of the line graph context.

The language model used is the 5-gram corpus from Google Books (Michel et al., 2011). The system selects all the 5-gram instances that were found to contain a synonym of the concept being expanded co-occurring with one of the words from the "concept context". The concept context is the set of head nouns that can appear in a sentence with the concept being expanded; in the example above, the concept context for "show" would be the terms "image", "graph", and "trend", since the possible contexts are the sentences: "The image shows a graph" and "The graph shows a trend". This set of lexical contexts is the same one used to seed the lexical expansion of concepts described above and originated from the most common terms used to express concepts in the experiment presented in (Greenbacker et al., 2011).

However, not considering the part-of-speech of the terms in the *n*-gram still allowed many terms that would not be applicable to replace the concept being expanded. The following example illustrates this phenomenon. For the term "volatile", the synonym "light" appears in the n-gram with the context word "trend". However, volatile is used in the context of the system as an adjective, a modifier of graph and trend. The way to fix this was to consider the part-of-speech of the context terms and the synonyms from the expansion when collecting synonyms from the language model. The Google Books corpus (Michel et al., 2011) has both POS tagged

and untagged 5-gram counts from digitized books. We used the POS tagged corpus for this. Without considering the part-of-speech of the grams, occurrences such as the ones presented below could not be avoided:

```
The_DET light_NOUN of_PREP the_DET trend_NOUN |count: 196
light_NOUN of_PREP the_DET recent_ADJ trend_NOUN |count: 79
throw_VERB light_NOUN on_PREP the_DET trend_NOUN |count: 40
```

In this example, the word **light**, which appears in the set of synonyms for the adjective **volatile** as shown in Figure 7-8, co-occurs in the corpus with the noun **trend** only as a noun as well. When considering the tagged adjective **light_ADJ**, the co-occurrence disappeared.

Figure 7-8: Thesaurus' synonyms for concept volatile used as a verb with a sense of "changeable".

The final set of synonyms that were disambiguated using this approach was still large and there were terms that were still inappropriate for the domain – although the subset acquired was considerably better than the initial set (few good synonyms were filtered out while a good number of inappropriate ones were). It was evident that

further filtering was required. As an alternative, we then decided to try a vector space model approach trained on Wikipedia (Wikipedia, 2004) data, which is available as a default corpus for training the word2vec tool available at (Mikolov, Chen, Corrado, & Dean, 2013). This approach is explained in the next section.

7.7.2.2 Using word vector space models to further filter appropriate synonyms

Effort has been made on using concept databases such as WordNet (Fellbaum, 1998) in order to disambiguate concepts. The work presented in (Klapaftis & Manandhar, 2005) combines the use of WordNet with search results from Google in order to allow the disambiguation of the word senses to be unsupervised. Word representation in vector spaces has shown to be a promising tool for acquiring terms' semantic knowledge. According to (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), experiments have shown that the word vectors capture many linguistic regularities.

As mentioned in the previous section, the set of synonyms filtered by using 5-gram language modeling yielded reasonable results but there was still room for improvement. We needed a secondary step that could help us further identify and collect synonyms that were applicable to the domain of line graphs. We then decided to use word vector space models. This technique builds vectors which represent the context of a term. The vector for the term "house", for example, has a higher count for the terms "big", "white", "spacious", than for the terms "hungry", "bag", and "sky". The vector is built by assigning co-occurrence counts to all the words in the language in question, and two terms can be compared on how similar they are in their contexts by measuring the similarity of their vectors. The idea is that two synonyms

ought to occur in the same linguistic; therefore, their word2vec scores should be very close.

When using the tool to assess its usefulness to the problem posed in this work, it was identified that the top ranked term was often its antonym (**rising** had **falling** at the top of the list of similar concepts).

By using the word2vec tool, the system was able to filter the set of synonyms collected from the language model step and further customize it to the line graph domain context. The reader might ask why both steps are needed in order to come up with the set of appropriate synonyms. As mentioned before, it was noticed that the language model alone was not sufficient since no threshold could be set in the system in order to consider a synonym for inclusion in the set. The reason being that any threshold eliminated the chances of good synonyms for the context of line graphs (volatility, for example), that were not as commonly used in the literature, from being added to the set. Using word vector representations alone, on the other hand, poses another challenge. The approach used by vector representation does not allow differentiation of a synonym from an antonym. The words "pretty" and "ugly" would have a very similar vector representation since they can be used within the same context. By collecting synonyms from a dictionary and starting the set of possible replacements from them, the antonyms were already filtered. By filtering co-occurrence present in Google N-grams (generated from digitized books), the noise is significantly decreased. One can then perform additional filtering by looking at the vector space models of the senses being disambiguated, which has good results for the line graph use case. The set of all final synonyms for each concept in our domain are presented in Appendix B. This combined approach proved to be a way of allowing a

system to create a customized *synset* of a domain by starting from a set of context words.

7.7.3 Grade level based lexicon creation

So far creating customized *synsets* for the line graph domains starts from a set of context words and expands the concepts by gathering all the synonyms assigned to the same part-of-speech as the word usage. Then it uses words from the small context to which the concept being expanded belongs (all of the possible uses) to look for their co-occurrence in a corpus by applying language modeling techniques. Last it employs the use of word vector space models in order to further filter the set of synonyms which semantically applied to the topic signature in question.

These steps enabled the system to come up with a set of terms that were appropriate lexical items for the line graph concepts needed for our summaries. Since the focus of the system is to generate text at different grade levels, a step to bin those terms based on their grade level appropriateness was also necessary. For any given concept, some of the lexical items may be rather simple and others might be considered more advanced.

In order to build grade level appropriate lexicons, the final set of synonyms disambiguated for the line graph domain was further divided into grade levels by checking for their lemma forms in the data previously used to learn text complexity feature measurements. From this step, each group of grade levels ended up with one or more terms that could describe the concepts used to generate descriptive summaries of line graphs. Since lexical choice can affect the final readability measurement of the generated text, the system randomly selects terms at the target reading level that will

represent concepts before starting the graph search explained earlier in this chapter. Evaluation results for the micro planning phase are presented in the next chapter.

7.8 Examples of Summaries Generated for Different Grade Levels

The following examples show the summaries generated for the line graph in Figure 7-9 for 4^{th} – 5^{th}, 6^{th} – 8^{th}, 9^{th} – 10^{th} and 11^{th} – college level.

4^{th} – 5^{th} summary:

There is an image. The image shows a line graph. The share of new homes sold before completion in percent is given by the graph. The graph consists of a changing trend composed of a rising trend from 1996 to 1999 followed by a stable trend through 2006. The graph is variable. The graph has the top value of 78.09 percent. The graph has the lowest value of 62.65 percent.

6^{th} – 8^{th} summary:

There is an image. The image reveals a line diagram which presents the share of new homes sold before completion in percent and consists of a changing trend composed of a rising trend from 1996 to 1999 followed by a stable trend through 2006. The diagram is variable. The diagram, which has the lowest value of 62.65 percent, has the highest value of 78.09 percent.

9^{th} – 10^{th} summary:

A volatile line diagram, which presents the share of new homes sold before completion in percent and consists of a changing trend composed of a rising trend from 1996 to 1999 followed by a stable trend through 2006, is shown by the drawing. The diagram, which has the lowest value of 62.65 percent, has the highest value of 78.09 percent.

11th – College summary:

A line graph, which presents the share of new homes sold before completion in percent and consists of a changing trend composed of a rising trend from 1996 to 1999 followed by a stable trend through 2006 and reveals some variability, is revealed by the image. The maximum value of 78.09 percent is reached by the graph which has the minimal value of 62.65 percent.

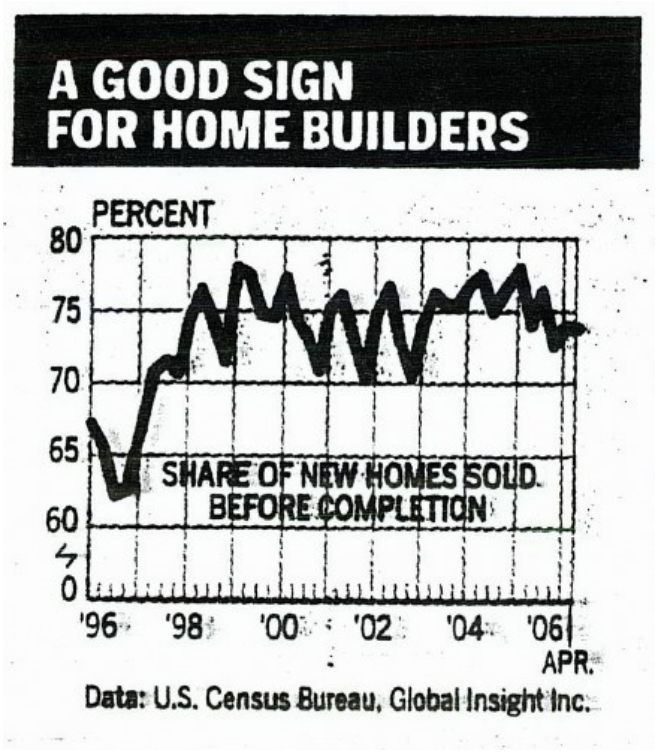


Figure 7-9: Example of graph extracted from online popular media.

7.9 Summary

This chapter described the micro planning phase of the SIGHT system. This phase aims to generate summaries of line graphs that vary in their readability levels by matching the readability level of the text surrounding the graph. The main steps for

enabling the system to adapt the generated text to different reading levels were described. The first step is concerned with understanding the aspects that make text complex. This was done by using decision tree learning to determine the set of features and their values that characterized each reading level. The second step is concerned with being able to measure them and map them to decisions made at the aggregation phase. The third step is to choose the set of relevant (to the context) and appropriate (to the reading level) lexical items that will be used to describe line graphs; here a multi-step process was used involving definition of base lexical items, synonym expansion, word sense disambiguation and target reading level filtering. The chapter details the approaches and methodologies used in this complex, but crucial phase of the NLG pipeline.

The next chapter describes all the evaluation experiments performed for this work. It describes both automatic and human judged assessments of the reading level and understandability of texts as perceived by participants at different reading levels. It also presents the evaluation performed with users with visual impairments.

Chapter 8

SYSTEM EVALUATION

This chapter presents the evaluations performed for assessing how successfully the system generates summaries at different reading levels and how useful these summaries are in delivering the high-level message to visually impaired users.

For evaluating the generation of summaries at different reading levels, four different aspects were considered. The first evaluation assessed how well the system was able to produce a summary that, according to automatic assessment tools, had a readability level that was close to the reading level of the text surrounding the graphic. The second evaluation was to allow the system to generate, for each line graph, all the possible reading level summaries and make sure they indeed differed from each other and matched the desired reading level. The third aspect evaluated was the perception of human readers, Mechanical Turkers who were presumably adults, about how these summaries could be ranked regarding complexity (in order to assess if people had the same "feeling" that one summary was indeed more complex than another). Mechanical Turkers are regular users who sign up to work on Human Intelligent Tasks (HITs) in the Amazon Mechanical Turk tool. And the last aspect was to evaluate whether the generated summaries were actually more appropriate for readers at the matching reading level than summaries that were far from it. For this last experiment, two sets of participants were recruited: 5th graders and freshmen college students. The task, explained in detail later in this chapter, assessed their preference for text

generated at a 4^{th} – 5^{th} grade reading level or at 11^{th} – college reading level and the reasons they chose one over the other.

The last section of this chapter describes the evaluation performed with visually impaired users that assessed the utility of the system in providing access to line graphs. Here we measure the ability of these participants to answer important questions about the graphic based on the generated summaries. This evaluation is presented in detail and it was one of the most challenging, given the difficulty of recruiting subjects for the experiment. Evaluation results and conclusions follow each of the experiments.

All the experiments performed had their protocols approved by the University of Delaware Human Subjects Review Board.

8.1 Evaluation of Summaries Generated at Different Grade Levels

The challenge with evaluating high-level summaries of line graphs is mainly the lack of a baseline or a comprehensive corpus of such summaries. Our evaluation task is even more challenging since we decided to generate these summaries at different reading levels in order to match the reading level of the surrounding text, as described in Chapter 7. So how do we even start evaluating these summaries?

The main idea behind dynamically choosing the target reading level is to make the summaries understandable to a reader of the text in which the summary appears regardless of the reading level of that text. With that in mind, some aspects need to be considered when evaluating such summaries: 1) Is the system able to, given a target reading level, generate a summary that is as close as possible to that target? 2) Can the system generate different summaries appropriate for different grade levels for each graph? 3) Do human readers indeed perceive texts being generated at different

reading levels by the system to be of different complexities? 4) Once the system is able to successfully generate summaries at different reading levels, it is necessary to evaluate whether what the system considers to be ideal for readers at a given reading level indeed is. So, evaluating how well summaries at the "right" reading level can help understanding and be preferred by readers at different reading levels is paramount.

8.1.1 Being able to generate summaries to match a given target reading level

The first step at assessing how well the system is doing when trying to generate summaries at the same reading level as the surrounding text of the graphic is to measure the reading level of that text. For that, the system uses the Style & Diction tool (Fsf, 2005) presented and discussed in Chapter 7. This tool provides readability measurements from a set of different metrics available in the field. Since these metrics do not always agree exactly on the readability level of a passage, we decided to assume a good metric to be the average of the ones which provide the grade level as their output. The metrics used to calculate this average were Flesch-Kincaid (Collins-Thompson & Callan, 2005), ARI (Kincaid et al., 1975), Coleman-Liau (Coleman & Liau, 1975), Fog index (Kincaid et al., 1975) and SMOG (Laughlin, 1969).

To determine the target reading level, the system takes as input the article in which the graph appeared. After running Style & Diction, the results are parsed in order to extract the measurements from the five metrics noted above. The average is calculated to identify the target reading level of the desired summary. In our first experiment, we generated the graphic summary at the target reading level and then measured how closely the generated summary matched the reading level of the article when assessed using the same tool.

For evaluating how well the system does on such a task, we used a set of 11 line graphs, which will be used throughout the evaluations of readability-aware generated text. For each one of the graphs, we ran the system five times, generating five slightly different summaries at the reading level identified for the article in which the graph appeared. These five summaries differ since, on each iteration of the system, the lexical choice randomly selects lexical items from the pool of appropriate options. Therefore, since the lexical items also affect the final reading level of the text, these five different runs had slightly different reading level measurements. Table 8-1 shows the results of this experiment. The grades used are represented by the ranges of grades used during system development. These ranges are 4^{th} – 5^{th}, 6^{th} – 8^{th}, 9^{th} – 10^{th}, and 11^{th} – College.

The results presented in Table 8-1 show that there was only one graph where the system generated summaries that were far from the target reading level. In this case, the generated summary had a lower reading level than the target. This graph, shown in Figure 8-1, had only one trend and few outstanding visual features (there was no extreme volatility or steepness of the trend, for example). Given that only a few propositions were selected for this graph, the graph search was unable to perform more grammatical modifications than it did, yielding a lower readability level than the target one.

Table 8-1: Results from matching the generated summaries reading level with the reading level of the surrounding text.

Graph	Target reading level bands	Summary reading level achieved (average of five runs)
L3	9th -10th grade	8.8
L6	6th - 8th grade	6.5
L17	6th - 8th grade	8.7
L18	6th - 8th grade	7.1
L21	11th - College grade	10.7
L23	6th -8th grade	8.2
L26	11th - College grade	10.9
L28	**9th - 10th grade**	**7.2**
L42	9th - 10th grade	9.5
L89	6th - 8th grade	7.3
L95	6th - 8th grade	7.3

8.1.2 Being able to generate summaries at distinguished reading levels for all graphs

This evaluation concentrated on the difference between the reading levels of the generated summaries when the system is configured to generate summaries at all grade levels for each of the 11 line graphs presented above. The results are presented in the next section.

Since each graph contains a different number of selected propositions and, therefore, different summary lengths, a way for evaluating its scalability across different graph types was to evaluate how well the system would perform when

configured to generate, for each graph, summaries at the four different grade bands available.

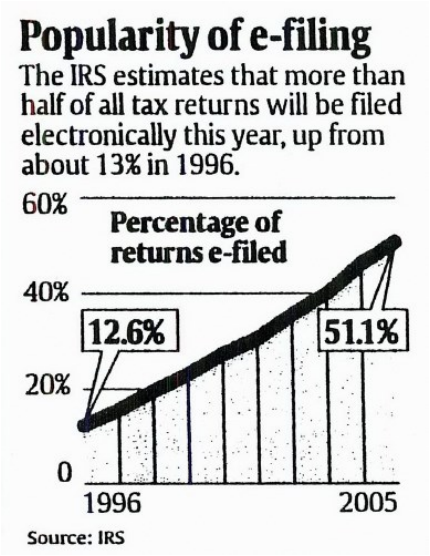


Figure 8-1: Graph named L28 used in the experiment.

Using the same rationale as in the previous experiment, we generated each summary five different times and averaged their readability measurements. The same set of metrics described in the previous section were considered for this phase.

Table 8-2: Experiment results for generating summaries for all graphs at all available grade level bands. Grade level bands marked with an * are the ones from the article in which the graph appeared.

Graph	**Grade level bands**	**Summary reading level achieved (average of five runs)**
L3	4^{th} – 5^{th} grade	6.1

Graph	Grade level bands	Summary reading level achieved (average of five runs)
	6th – 8th grade	7.8
	9th – 10th grade*	9.4
	11th – College grade	10.19
L6	4th – 5th grade	4.7
	6th – 8th grade*	6.1
	9th – 10th grade	11.6
	11th – College grade	12.2
L17	4th – 5th grade	6.5
	6th – 8th grade*	8.6
	9th – 10th grade	**12.6**
	11th – College grade	**12.6**
L18	4th – 5th grade	5.4
	6th – 8th grade*	7.3
	9th – 10th grade	9.9
	11th – College grade	10.5
L21	4th – 5th grade	5.2
	6th – 8th grade	7.1
	9th – 10th grade	**10.8**
	11th – College grade*	**10.8**
L23	4th – 5th grade	6.5
	6th – 8th grade*	7.8
	9th – 10th grade	10.9

Graph	Grade level bands	Summary reading level achieved (average of five runs)
	11th – College grade	11.6
L26	4th – 5th grade	5.2
	6th – 8th grade	6.8
	9th – 10th grade	**10.2**
	11th – College grade*	**10.2**
L28	4th – 5th grade	5.6
	6th – 8th grade	7.7
	9th – 10th grade*	11.2
	11th – College grade	11.6
L42	4th – 5th grade	3.9
	6th – 8th grade	6.1
	9th – 10th grade*	8.9
	11th – College grade	9.8
L89	4th – 5th grade	4.4
	6th – 8th grade*	7.3
	9th – 10th grade	**11.7**
	11th – College grade	**11.7**
L95	4th – 5th grade	5.0
	6th – 8th grade*	7.4
	9th – 10th grade	**10.8**
	11th – College grade	**10.8**

The overall grade level average for 4th – 5th grade is 5.2; for 6th – 8th it is 7.29; for 9th – 10th it is 10.73; and for 11th – College it is 11.08.

Even though the reading levels are constantly increasing, for some line graph summaries, the reading levels remained the same for the last two grade bands (9th – 10th and 11th – College). This result (bolded in Table 8-2) can be explained by the fact that such graphs had few propositions selected by the content determination phase, so there were fewer options for aggregation and grammatical combinations. By comparing the summaries for graphs L26 and L89 below, we can notice that both had about the same number of propositions but graph L26 could not aggregate the last sentence (*The minimum value comes about on 10/2005.*) due to the constraint on embedded relative clauses. Having that isolated sentence probably lowered the measures making L26 have a slightly lower reading level than 11 for the 11th – college grade level band. The heuristic probably decided to aggregate all the possible propositions for graph L89 since not aggregating them could have resulted in a grade level far lower than desired.

(Summary for graph L89)

A line graph, which presents the share of new homes sold before completion in percent consisting of a changing trend composed of a rising trend from 1996 to 1999 followed by a stable trend through 2006, is shown by the image. The top value of 78.09 percent is reached by the graph which has the lowest value of 62.65 percent.

(Summary for graph L26)

A line graph, which presents the number of consumer confidence and consists of a trend that changes and that starts to reverse at the end composed of a falling trend from 5/2005 to 10/2005 followed by a rising trend through 4/2006, is revealed

by the drawing. The highest value of 111.07, which comes about on 4/2006, is reached by the graph which has the minimum value of 87.09. The minimum value comes about on 10/2005.

These results show that the system successfully produces summaries that are close to the target reading levels that were given as input to it. Some exceptions exist where the reading level of the produced text is out of the expected range, however the majority of the time the summaries are within the range. Thus, we conclude that the methodology is fairly accurate.

8.1.3 Being able to match human readers' perception of text complexity

Even though we confirmed that the system was doing a good job at generating text at different reading levels according to automated tools, we still did not know if these different reading levels were also perceived by humans to be different from each other and, even more, if their ordering of simplest to most complex matched human judgement.

To evaluate whether the different readability levels generated by the system were also perceived the same way by humans, an evaluation with human subjects through crowd-sourcing was performed. The main idea was to be able to measure how well human readers would "rank" or "order" line graph summaries based on their readability levels and compare that to the ordering the system was aiming for when generating them.

With that in mind, an experiment using Amazon Mechanical Turk was performed in order to measure how well the system was doing in generating text whose reading level was indeed ranked from least to most complex by humans. Appendix C shows the Human Intelligent Task provided to the *turkers*. The following

subsections detail the task *turkers* (crowd-sourcing workers who log in to execute a task) were asked to undertake, the analysis of the data done in collaboration with a team of statisticians at IBM Watson Analytics research team (Chu, Y., Li, L., Shyr, J. Y., personal communication, September 30, 2015 - October 15, 2015) and the evaluation of the results.

8.1.3.1 Defining the HIT (Human Intelligence Task)

The main idea behind the task (called a HIT in Mechanical Turk) was to identify whether human readers would order the summaries of a line graph from simplest to most complex in the same order the system generated them. For the tasks, 10 graphs[9] were used. A control task was used to make sure the *turkers* understood the task, and only the answers from *turkers* who got the control question right were considered. Below is the experiment task the same way it was shown to the *turkers*. This same configuration was used for all graphs, only changing the order in which the summaries at different reading levels were presented from one graph to another. Each graph could be answered by nine different *turkers* and each *turker* could only answer one HIT per graph but an individual *turker* could perform the HIT associated with one or more different graphs.

---

[9] One of the graphs, L17 was not included in this experiment because it was very similar to another graph which was already part of the experiment. We wanted to get representative examples of graphs with varying salient visual features and intended messages to be able to evaluate how these would affect the results.

8.1.3.2 Analyzing the data collected from the experiment

After the experiment was run, a total of 90 HITs were undertaken (10 graphs, 9 *turkers* per graph). Each one of these HITs produced an ordering which corresponded to the grade level the *turkers* believed the summaries belonged to. It is important to notice that the instructions on the task clearly stated that they could only choose each grade level once. This requirement was posed in order to somehow force them to order the summaries as best they could, instead of just repeating grade levels for the summaries that could be similar in readability. This requirement posed a challenge, however, since choosing the wrong grade level for one summary forcibly led them to choose a wrong grade level for at least one other summary.

The challenge that came with the added constraint regarding associating a unique grade level to each summary was more accentuated when evaluating the results. Since this would affect the results, just counting how many associations were made correctly did not seem to be the right approach. We needed, instead, to determine how close the results were to the system's ordering and how well the summaries were categorized regarding their grade levels. If we could at least see if a summary written at 4^{th} grade level was classified as a 7^{th} grade level instead of College level, we would like this to count as a positive result.

The Pearson correlation coefficient did not seem to be the right approach since it would not allow us to take into account the closeness to the desired output noted above. Spearman correlation produced the same results as Pearson when applied to our data. This might be because the set of values were already the summaries' ranks, so when applying both formulas the same results were obtained.

We decided then to consult with a SME (subject matter expert) and ask what they thought a good statistical formula or approach would be for this data in order to

measure the aspects we were looking for. From correspondence with three statisticians from the Watson Analytics team at IBM (Chu, Y., Li, L., Shyr, J. Y., personal communication, September 30, 2015 - October 15, 2015), this was the initial response from the statistics team:

"The question you have isn't something that immediately fits into a standard protocol we usually handle here. After discussing with some co-workers, I wrote a document about how we would tackle it in two ways (because we don't know if you care the position or ordering more) based on the nine records you had. Please take a look and see if they make sense. Please also let me know if you have any questions."

After some clarifying email and analysis of the document, the ordering approach was indeed the one we were looking for. The approach is explained in the next subsection and the results of applying it to our data follows.

Additionally, we evaluated how good the ordering provided by the *turkers* was by calculating the nDCG (normalized Cumulative Discounted Gain) scores for each graph. This score is used in Information Retrieval in order to assess how well a search engine returned results for a query based on relevance labels associated with the search results. In this scenario, the different reading level texts were associated to relevance orderings and the expected DCG was used to normalize the DCG of the provided orderings. In Information Retrieval, however, search results can be judged as irrelevant to the search query at all (usually having a relevance label of 0). That is different from the scenario of this work since all the documents had some relevance which represented the ordering of the summaries grade levels. The nDCG results were higher than the percentages found by applying the pairwise relationship approach explained next. The results of the nDCG analysis follows the pairwise relationship

and a table on which the formula was used to calculate the results is shown for graph L3. nDCG is a good measure to compare results across queries (or graphs, in this scenario).

8.1.3.3 The pairwise relationship approach

There are six distinct pairwise relationships among the four outcomes in the data containing the results from the tasks. The method is to count the number of times the orderings among the pairs are in the desired direction for each of the six pairs and combine them over all pairs.

For an example where the expected ordering is **3 4 1 2**, the six pairwise relationships are: $\mathbf{3 \succ 4;\ 3 \succ 1;\ 3 \succ 2;\ 4 \succ 1;\ 4 \succ 2;\ 1 \succ 2}$, where $\succ$ represents that 3 should come before 4 in the assignment of grade levels. For nine possible orderings that could represent a response from nine different participants, the following table shows how the pairwise relationship measurement would be obtained.

Table 8-3: Table with an example of the Pairwise relationship approach proposed by the statistics team from Watson Analytics (count indicates the number of correct pairwise relationships).

	Pairwise relationship						
Responses	**$3 \succ 4$**	**$3 \succ 1$**	**$3 \succ 2$**	**$4 \succ 1$**	**$4 \succ 2$**	**$1 \succ 2$**	**count**
(1) 4 3 1 2		x	x	x	x	x	5
(2) 3 4 1 2	x	x	x	x	x	x	6
(3) 4 3 1 2		x	x	x	x	x	5
(4) 3 2 1 4	x	x	x				3

	Pairwise relationship						
Responses	**3 ≻ 4**	**3 ≻ 1**	**3 ≻ 2**	**4 ≻ 1**	**4 ≻ 2**	**1 ≻ 2**	**count**
(5) 4 3 2 1		x	x	x	x		4
(6) 3 4 1 2	x	x	x	x	x	x	6
(7) 4 1 3 2			x	x	x	x	4
(8) 4 3 1 2		x	x	x	x	x	5
(9) 2 1 3 4	x						1
count	4	7	8	7	7	6	39
Prob (%)	**44.4%**	**77.8%**	**88.9%**	**77.8%**	**77.8%**	**66.7%**	**72.2%**

The result of 72.2% can then be used to assess how good the overall results are (100% being the perfect ordering). This overall measure is based on an equal weight for each pair since no pair is more important than another. If it was the case some pair was more important than others, then different weights could be given and a weighted average could be used.

Appendix D contains all of the valid responses and the pairwise relationship results for each response. It also shows the calculation for finding the nDCG for graph L3, as an example.

8.1.3.4 Results using the pairwise relationship approach

Looking at the overall results, available in Appendix D, a total number of 348 pairwise relationships were valid. From these, 252 were correct. This yielded an overall average of 72%, which represents how well the system's text complexity rankings compared with human judgement.

When examining the graphs individually, we notice a big discrepancy in the results. Half of the graphs had their correct pairwise relationship count above 80%, 3 out of 10 graphs were between 65 – 70% and 2 out of 10 were between 45 – 55%. By examining the graph images for these sets there was no conclusion about what could have caused such a discrepancy on the set of results. Characteristics that were examined were the set of outstanding visual features of the graphics and the number of propositions selected by the content determination algorithm.

A possible explanation for such discrepancy could be the fact that the lexical items that belonged to lower grade levels could also be used in higher grade levels. In cases where the summaries for different grades used similar lexical items, it is possible that the participants judged a summary to belong to a lower grade level than it was classified by the system.

8.1.3.5 Results using nDCG score

Table 8-4: Results of applying nDCG to results from *turkers*.

Graph	Average nDCG
L3	0.9598
L6	0.9860
L18	0.8975
L21	0.8893
L23	0.9451
L26	0.9752
L28	0.9888

L42	0.9365
L89	0.9851
L95	0.9798

The results of applying nDCG are higher than the ones gotten from the pairwise relationship approach. Although the nDCG score is a useful metric for evaluating relevance ranking, it might not be the most appropriate metric for evaluating the results of the task described in Section 8.1.3.2 since it penalizes top ranked results more and we would like to penalize misplaced assigned summary grades according to their distance from the target reading level.

8.1.4 Being able to generate text at reading levels that are indeed appropriate to readers at different reading levels

The last evaluation was to assess if readers at different reading levels indeed prefer summaries generated at their reading levels. Two groups of participants were recruited for this experiment: (1) students from a fifth grade elementary school in the Austin, Texas area and (2) undergraduate students in an introductory CS course at a University.

Participants were presented with two summaries for each of nine different graphs. For this experiment, participants had access to the graph images. Since one of the graphs proved to be misleading in a previous experiment regarding the description of its intended message (the segmentation did not recognize a slight drop in a rising trend) we decided to remove that graph.

One of the summaries was generated for the 4th – 5th grade level and the other for the 11th – College grade level. The participants were asked to select the summary they liked the best and to provide comments about what they did not like in either

summary. Instructions for the experiment were as follows (the example below was used for college students. For the fifth graders the instructions were written at their reading level):

1. *This package contains a set of trials, each consisting of a line graph and initial summaries that convey the high-level content of the graph. Please take as long as you need for each trial. You will have about 45 minutes. Once the 45 minutes are up, please complete the trial you are working on and stop.*
2. *The trails are designed to evaluate the generation of summaries that are intended to possess the same text complexity level of the article's text. This initiative aims to leverage the user's experience when reading an automatically generated summary of a graph.*
3. *Each trial is composed of a graphic and two automatically generated summaries. You should read the graphic and the summaries present on it. Summaries from a given graphic have the exact same content but their text complexity changes based on different reading levels. You will be asked to choose the initial summary you preferred and provide us with details on why you prefer one over the other.*

We hope you enjoy participating in this experiment and we profoundly appreciate your collaboration.
SIGHT System Team.

Then the participants were provided with a line graph image and two summaries, one generated at the 4th – 5th grade level and another generated at the 11th – college grade level. They were then asked the following questions:

College students:

Please answer the following questions:

1) Which summary did you prefer?
2) Why did you prefer this summary?

3) Please tell us what you DID NOT like on either summary. You can also circle the passages that you think should be written differently and let us know why.

5th graders:

Questions:

1) Which summary did you like more? _____

2) Why did you like that summary better? (space for answer was provided)

3) Do you think one of the summaries would be easier for a classmate of yours to understand? _____ Yes _____ No

If yes, which summary do you think your classmate would prefer? _____

4) If there is anything in either summary that you did not like, please circle or underline the part(s) on the summaries in the previous page.

5) If you circled anything, please tell us why you did not like those parts on:

SUMMARY 1 (space for answer was provided)

SUMMARY 2 (space for answer was provided)

8.1.4.1 Experiment performed with 5th graders

Sixteen 5th graders were recruited to take part in this experiment. Each trial package contained five different graphs, randomly selected; each 5th grader had 30 minutes to work on as many graphs as they could. The children received a small school supply kit for their participation. Table 8-5 shows the results of the experiment performed with 5th graders. These results show that the majority of the children preferred the summary generated at the 4th – 5th grade level. Besides stating which

summary they preferred, they were asked to also provide the reason for the choice. Some of the comments from the children who chose the 4th – 5th grade level summary were:

"it has more information and it stays on topic. It explains about it and it has more details." (participant A)

"It was informal but still easy to understand. Also the words used caught my attention" (participant B)

"It didn't give big confusing words" (participant C)

"I prefer active voice" (participant A)

"it explained better than the other" (participant D)

"It was easier to understand. The summary also uses descriptive words." (participant E)

"Because this summary made more sense than the other" (participant F)

"Because I think a kid like me would understand it better" (participant G)

Some of the comments from the kids who chose the summary generated at the 11th – College grade level were:

"Because it explains more about the graph" (participant H)

"I did not like how direction was used in the summary" (participant I)

"it gives more info and uses bigger words" (participant J)

"It gives a lot of information about the barrels. It tells you the dates of the barrels." (participant K)

Note the comment of participant J (who preferred bigger words). This comment reaffirms the decision in SIGHT to generate at the level of the article being read. Despite this student being in 5th grade, s/he seems to prefer reading at a higher grade level. Thus, each reader may prefer material at different reading levels. Assuming all the children would prefer to read text at the school grade level they are currently at is an erroneous assumption. Some children might read more frequently than others, increasing their ability to absorb more complex text, while other read at below their grade level.

The overall results from the experiment show that 78.08% of the time the children preferred the summary that was generated at the 4th – 5th grade level.

Table 8-5: Results from reading level experiment with 5th graders.

Line graph	**Preferred summary generated at the 4th – 5th grade level**	**Preferred summary generated at the 11th – College grade level**
L6	9	3
L17	10	1
L18	1	0
L21	6	3
L26	5	2
L28	8	2
L42	7	0
L89	7	1

Line graph	Preferred summary generated at the 4th – 5th grade level	Preferred summary generated at the 11th – College grade level
L95	4	4
Total	**57**	**16**

8.1.4.2 Experiment performed with freshmen College students

The experiment with freshmen College students was performed in the same fashion as the one with 5th graders. They were presented nine different graphs that were randomly organized in each individual trial package. The college students were given 45 minutes to work on as many graphs as they could from a package of nine graphs. The students were recruited from three different Summer classes and they were all offered a small amount of extra points for their participation. 34 students took part in the experiment.

As shown in Table 8-6, there were a total of 163 responses. The percentage of students who chose the summaries generated at the 11th – College grade level was 70.55%.

Table 8-6: Results from reading level experiment with freshmen College students.

Line graph	Preferred summary generated at the 4th – 5th grade level	Preferred summary generated at the 11th – College grade level
L6	5	13
L17	6	14
L18	4	15

Line graph	Preferred summary generated at the 4th – 5th grade level	Preferred summary generated at the 11th – College grade level
L21	6	16
L26	5	14
L28	5	15
L42	5	10
L89	6	10
L95	6	8
Total	**48**	**115**

College students were also asked to highlight the fragments in either summary that they did not like and provide reasons why they didn't. Some of the comments by participants who chose the summary generated at the 4th – 5th grade level were:

"Presents info in an order which is easier to understand" (participant A)

"More concise and understandable; could combine some sentences" (participant B)

"Easier to follow along with the shorter sentences" (participant C)

"More coherent and keeps each sentence to one idea" (participant D)

Some comments about the reasons why the participants chose the summary generated at the 11th – College grade level were:

"More accurate by using less words" (participant E)

"More fluid, continuous conveyance of information" (participant F)

"More flow, allows for the reader to just read that and understand clearly" (participant G)

"It is more clear to understand than summary 1 because of the way the sentences are presented and where the words are placed" (participant H)

"Better readability, sentence structure and flow; the sentences fit well together and therefore made the passage easy to read and gather information from" (participant I)

"Grammar used is better" (participant J)

"It gave all the information in a non overwhelming way" (participant K)

From the results obtained from the trials performed with 5th graders and the ones obtained from the trials performed with college students, we can conclude that indeed the texts generated at their reading level were better accepted by the majority of the participants, in both cases. The exceptions confirm that not every reader in a given grade is at the same reading level, which reinforces the choice of assessing the target reading level by looking at the article in which the graph appears, since presumably they feel comfortable with the reading level used by the venue. Table 8-7 shows that the results are statistically significant given p = 3.67816E-12 calculated using the chi-squared test.

Table 8-7: Statistical significance results.

	5th graders	College students	Total	Prob
5th grader text	57	48	105	0.44
College text	16	115	131	0.56
Total	73	163	236	

8.1.4.3 Lexicalization analysis in the context of the previous two experiments

Another aspect evaluated was word choice. Remember that the previous experiments, performed with 5th graders and college students, asked the participants to underline anything that they did not like in the summaries they read. This included the word choice in the summaries. The participants indeed pointed out some words that they would rather replace.

In the responses provided by 5th graders, there were seven occurrences of word complaints, where they complained about the use of the adjective *jagged* for describing volatility and for the word *direction*, referring to a trend.

27 trials had complaints about word choice in the set of responses from the freshmen College students. Some examples were regarding the opposite concepts (*maximum* used with *lowest* and *top* used with *minimum* in the same summary). This occurs because the lexical items are selected in isolation and no coordination of lexicalization for opposite concepts is employed. This is an interesting problem and it is mentioned in Chapter 9 in the context of future research.

Some complained that the word *top* and *highest* should not be used, and that *maximum* should be used instead. The lexical item *early* was included in the set of synonyms describing the initial concept of a trend or graph. Some participants complained about the use of this word.

Some comments were:

"summary should use first and final instead of early and last" (participant A)

"use initial instead of early" (participant B)

Overall, for the 5th graders, there were wording complaints 9.5% of the time while for freshmen College students the number of complaints went up to 16.5%. It is important to notice that no agreement was found when counting the number of total word choice complaints (there were specific complaints that were made by only one of the participants throughout the responses).

8.1.4.4 Conclusions on the experiments

For the freshmen college students, the fact that almost 30% of the subjects chose the summary generated at the 4th – 5th grade level, even though they were all at the same grade level, was expected. Similarly, around 22% of the 5th graders chose the summary that was generated at the 11th – College grade level.

These results show that reading preferences may vary even among people from the same age/grade level, as noticed with the group of 5th graders. Since there were subjects who preferred simple text over complex text, we can assume that reading skills can vary even within a grade level group. Our contention is that readers who prefer simple text would read venues that use simple text structure and syntax. Instead

of assessing or asking the user their grade level, our approach provides more chances of being successful at producing text that will be appropriate to each user. From the experiments performed, we conclude that pursuing the generation of natural language text that fits the reading level of the surrounding text is promising.

8.2 Evaluation of Summaries with Users with Visual Impairments

For the purpose of evaluating the usefulness of the generated summaries, we performed a task-based evaluation with people with visual impairments. We decided to evaluate the system by testing if users with visual impairments who had access to the summaries would be able to answer important questions regarding the high-level knowledge conveyed by the graphic, and to compare their performance to that of sighted users viewing the graphic in answering the same questions.

The experiment was composed of three phases. The first phase was concerned with the collection of questions to be asked of the participants. It was important to collect an unbiased set of questions (i.e., not influenced by the thoughts of our project team concerning what is most important), and that the questions be something that could be answered with the graph and not require domain reasoning or information beyond the graphic. To make sure of this, the first phase was subdivided into four sub-phases: questions collection, question filtering, minimum agreement assessment and rewording a question, if needed. For the first sub-phase, we asked sighted users to provide us with questions that a person would be able to answer by just glancing at the graphic because the goal of the system is to initially provide a summary that will suffice for users who are reading a multimodal document containing graphics but who are not interested in analyzing the graphic in detail. The second sub-phase was the filtering of relevant questions. For this sub-phase, we asked another participant, who

did not take part in the first sub-phase, to filter the questions such that questions that required world knowledge or inference, and those where answering required a careful examination of the graphic, would not be included. The participant had knowledge about Natural Language Processing but not about the project itself.

After the questions were filtered, the set of questions to be used was restricted so that a question would only be included in the second phase if at least two of the participants posited that question. In this way we hoped to ask the most important questions. From the resulting set of questions, we could find questions that were worded differently but meant the same thing. Since some questions were asked in a clearer way than others, we asked the participant who filtered them in the first sub-phase to choose a clear way to state the question.

The second phase was the evaluation with visually impaired readers, where we provided them with the summaries and asked the questions collected during the first phase. The last phase was the collection of control answers. Sighted participants were recruited and provided with the graphic images and the same questions that were asked of the blind users. All the phases are described in detail next.

8.2.1 Phase 1: Collection

8.2.1.1 Collecting questions from sighted users

For this phase we recruited freshmen college students from various majors to provide us with questions that a person would be able to answer by just glancing at the graphic. By asking subjects to provide us with questions, we wanted to make sure that the questions were not biased by the system designers and that they reflected what users would actually ask.

Thirty-four students participated in this phase. This task was interspersed with another task, which provided data for another project. Since the second task is not relevant here, instructions for it were suppressed. The instructions for this phase of the task were the following:

> *1. This package contains a set of trials, each consisting of a graph and a task. For each graph you will be asked to either: A) Answer questions related to it. B) Provide questions about it.*
>
> *2. For the graphs which we ask you to provide questions about, we ask you to think of a question that a person would be able to answer by just taking a quick look at it. The question should be able to be answered by the high level knowledge the graphic conveys (without the need for calculations or detailed examination).*

For the trials in which we asked participants to provide us with questions, we gave them a line graph and asked them to provide us with at least two questions. Eleven different line graphs were used in this phase. Figure 8-2, Figure 8-3, and Figure 8-4 are three examples of graphics that were used. The graphics conveyed different intended messages and had different sets of visual features. A total of 216 questions were collected from 34 participants, an average of about 19 questions per graphic.

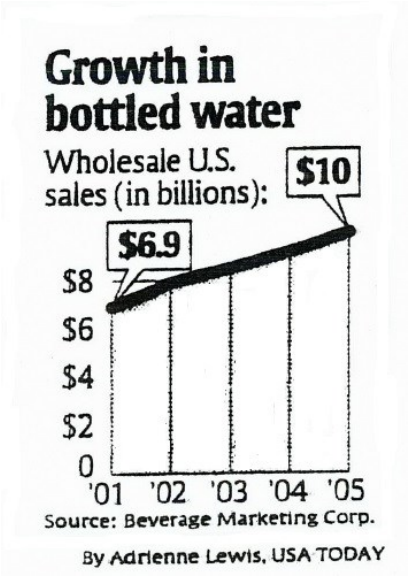


Figure 8-2: Example of a line graph showing a single rising trend used in the first phase of the experiment.

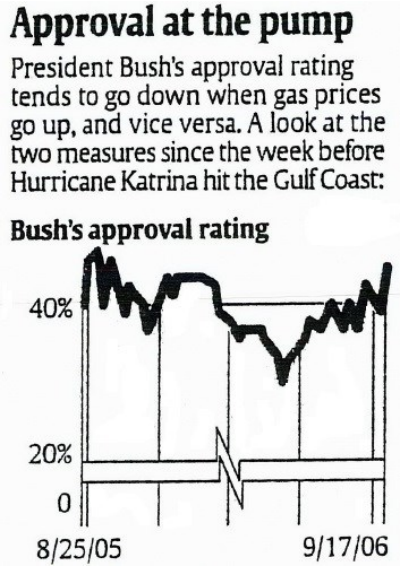


Figure 8-3: Example of a line graph showing a changing trend and some volatility used in the first phase of the experiment.

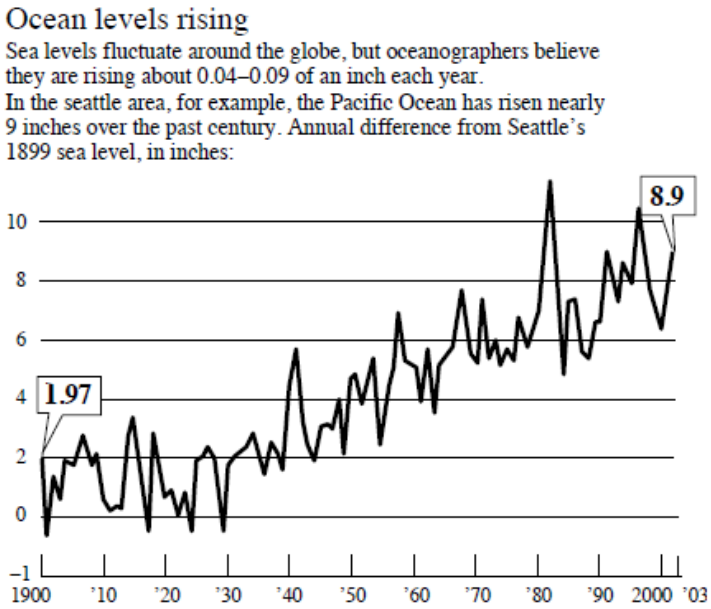


Figure 8-4: Example of a line graph showing a changing trend and strong volatility used in the first phase of the experiment.

8.2.1.2 Question filtering

A native English speaker participant with knowledge of Natural Language Processing was recruited for the second sub-phase, which was concerned with filtering questions that were appropriate to be asked. The participant was asked to filter out any questions which required world knowledge (the question could not be answered from the graph alone), careful examination of the graphic or complex calculations.

The first exclusion criterion was the need for world knowledge in order to answer the questions. Examples of questions that were omitted from the next phase of the experiment are:

1. **Why has there been growth in bottled water?** (Referring to the line graph shown in Figure 8-2);

2. **How does this inverse relationship impact society?** (Referring to the line graph shown in Figure 8-3);
3. **Do you think this percent change is similar to other sea levels in other parts of the world?** (Referring to the line graph shown in Figure 8-4).

Since such questions cannot be answered by just having access to the information conveyed in the graphic, they were eliminated by the participant performing the filtering.

Examples of questions that were excluded based on the second exclusion criterion (required a detailed examination of the graphic) are:

1. **What was the amount spent on bottled water in 2004?** (Referring to the line graph shown in Figure 8-2);
2. **What is the lowest percentage of his approval rating?** (Referring to the line graph shown in Figure 8-3);
3. **What year had the highest sea level, in inches, in Seattle?** (Referring to the line graph shown in Figure 8-4).

The last exclusion criterion was the need for a complex calculation in order to answer the question. It should be noted that questions that still required some quick calculations were left in by the participant performing the filtering. Examples of questions that required calculation but were still left in are:

1. **From 01 to 05 bottled water sales have grown by how much?** (Referring to the line graph shown in Figure 8-2);
2. **From 1900 to 2003, what is the total difference between sea levels in inches?** (Referring to the line graph shown in Figure 8-4).

And an example of a question that required complex calculation and was eliminated is:

1. **What was the percentage increase between sea levels from 1900 to 2003?** (Referring to the line graph shown in Figure 8-4).

8.2.1.3 Choosing/Rewording unclear questions

After filtering the initial set of questions, a total of 125 questions that were considered appropriate by the participant were grouped by meaning. Questions that were worded differently but meant the same thing were grouped together so that we could assess the agreement between participants about which questions they thought reflected the knowledge one could acquire from the graphic. Only questions that had been provided by at least two participants were used.

Since some questions (in the same group of meaning) were worded better than others and therefore were clearer, we asked the participant who filtered the question to either choose one question from each meaning group (the one that was clearer) or, if none was clear enough, we asked the participant to rewrite it such that the person reading the questions would not need access to the graphic in order to understand it. This was necessary because there were some questions that made references to the graphic or to the context of the previous question they provided, making the question unclear if standing alone. After filtering and choosing/rewriting the appropriate questions, we had a total of twenty-one questions about nine different graphics. These questions were used for the following phases of the experiment: assessing the usefulness of the summaries for visually impaired users and collecting control answers from sighted users who would answer the questions by viewing the graphic image.

8.2.2 Phase 2: Evaluation of the summaries with visually impaired users

For this phase we recruited four blind users with the help of the Delaware Association for the Blind. Given a line graph, the participants would have access to the summary generated by the system and would be asked one to three questions about the graphic. Navigation instructions were provided before the participants started listening to the task instructions.

Demographic information indicated that all four participants had been using the Internet for reading news for more than seven years and all of them use screen readers as their main reading tool (options were braille, screen magnifier, screen reader, or other). All of them had some college education (two of them had some graduate education, one of them had a Master's degree) and all of them were native English speakers.

For this phase we collected answers for all 36 trials. The participants had up to 45 minutes to perform their task of answering the questions. Instructions provided to the participants were:

Navigation instructions: At the end of each paragraph, the screen reader will pause. Please use the key combination Control plus down arrow to move to the next paragraph, the combination Control plus up arrow to move to the previous paragraph, and the combination Control plus right arrow to read the current paragraph.

SIGHT System Line graph summary evaluation - Experiment Instructions

First you will be asked to answer a small survey that collects demographic information. Then you will be presented with a set of trials. Each trial contains a summary of an informational graphic from some popular media available online (newspaper or magazine). For each trial, you will be presented with a summary of a

graphic and then asked to answer one to three questions based on that summary. The summary may or may not provide all of the information needed to answer the question, please just answer as best as you can. You may not understand some questions. At the end of each trial you will have a space to comment. Please let us know anything about the summary you would change or add. The question might ask you to calculate something. You may use a calculator if you wish or just give an approximate answer. You will also be able to give us comments on the questions at the end of the trial. Our intention is to measure how well the summary delivered the information that you needed to answer the questions. You can listen to each summary as many times as you wish before proceeding to answer the questions related to that summary. After answering the questions for a summary you will have the space for comments. At the end, you will be asked some questions regarding your overall experience and any additional comments you may have to help us improve the system. You have the choice of typing your answers and comments yourself, or having your spoken responses recorded. If you choose to type your answers and comments, please do so after each question when the screen reader reads "Empty Paragraph" or "Edit". In order to move the cursor to the empty paragraph and be able to edit the field, please press the key combination Control plus Space bar.

The results and comments provided by the participants in this phase were analyzed by comparing with the control answers provided by sighted users during phase 3 of the experiment, described in the next section.

8.2.3 Collection of control answers from sighted users

The focus of this phase was to assess whether users would be able to correctly answer the question viewing the graphic for baseline purposes. We wanted to know if

sighted users could answer the question by glancing at the graphic, but allowed them a more careful look at the graphic if they wanted and we recorded this information along with the correctness of their response. The results are also presented for both cases: percentage of correct answers from sighted participants by just glancing the graph and percentage of correct answers from sighted participants after being able to examine the graph carefully.

For this phase, 24 freshmen college students from various majors were recruited. The participants for this phase could not have participated in any of the previous phases of the experiment. As in the first phase, a different task was mixed amongst the trials in order to avoid task fatigue. The instructions for this phase were:

1. Same from Phase 1 (see page 164).

2. *For the graphics about which we ask you to answer questions, please study the graphic such that you may be able to answer the questions about it. You will then be asked to answer one to three questions about the graphic. You may not understand some questions or feel that they cannot be easily answered with the graphic. At the end of each question, you will have a space to make comments on it. Please let us know anything about the question you could not understand. The question might ask you to calculate something. You may use a calculator, if you wish, or just give an approximate answer. ATTENTION! While answering a question about a graphic you can go back to the image or just provide the closest answer you can remember. If you choose to go back, please let us know you did so in the space designated for it. In the cases where you do go back to the graphic, we ask you to estimate how difficult the question was to answer even with the graphic.*

Even though participants were told they could go back to the graph if they needed to, we were able to see from the results, described in the next section, that most of the questions did not cause them to go back to the image; the exceptions were the questions which required some simple calculation.

8.2.4 Evaluation results

As described in the previous section, the four participants with visual impairments received a package with nine trials. Each trial had a line graph summary and one to three questions that they were asked. For the collection of control answers, 24 sighted participants received a package with the same number of trials and the same graphics. The set of graphics contained 3 graphics that overlapped with the graphics used to identify features (described in Chapter 4). In order to make a clear distinction of the results, the graphics which overlapped will be signaled in the table showing the final results.

Table 8-8, Table 8-9 and Table 8-10 show the total number of questions answered by graph (since participants could stop before finishing the whole package), the number of those answers that were wrong or incomplete, and the percentage of correct answers. There is also a distinction made in the result presentation when considering the control answers from sighted participants when they went back to carefully analyze the graphic before answering the questions and when they answered by just glancing at the graphic (the main purpose of the experiment, since the summaries try to capture this type of information). The graphics which were also used in the experiment to identify features are marked with an * close to their codes.

Analyzing these numbers, we notice that line graphs L17 (Figure 8-5) and L26 (Figure 8-6) have considerably lower scores in the correct answers column. Although

the sighted participants did better with these two line graphs, we note that they often went back to the image to search for the answers (16 out of 27 went back to the image for the line graph L17 and 9 out of 15 went back to the image for the line graph L26). The questions, followed by the answers, for this line graph were:

1. Question 1: **How many pension plans were defined in 1985?** Answer: 114,396
2. Question 2: **How many more pension plans were defined in 1985 than in 2004?** Answer: 83,158
3. Question 3: **What is the percent decline of pension plans from 1985 to 2004?** Answer: Approximately 73%

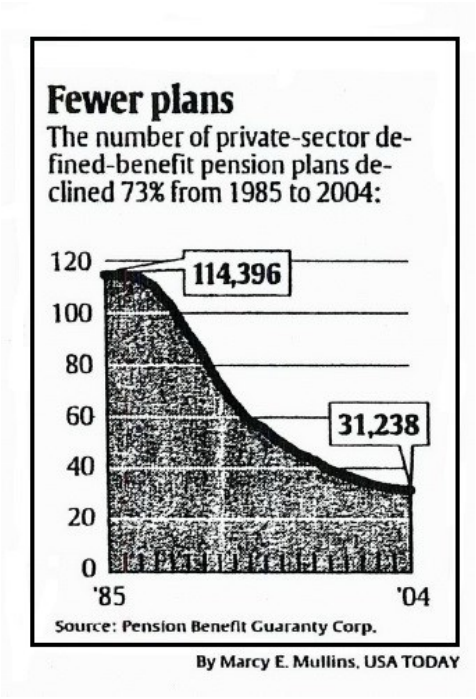


Figure 8-5: Example of a line graph used in the experiment (graph L17).

Summary generated for the graph in Figure 8-5:

The image shows a line graph, which presents the number of private-sector defined benefit pension plans. The line graph shows a falling trend from 1985 to 2004. The falling trend has a starting value of 114,396 and an ending value of 31,238.

Summary generated for the graph in Figure 8-6:

The image shows a slightly volatile line graph, which presents the number of consumer confidence in addition to conveying a trend that changes and that starts to reverse at the end that consists of a falling trend from 5/2005 to 10/2005, followed by a rising trend until 4/2006, then a falling trend through 5/2006. The first segment is the falling trend that has starting value of 103.1. The second segment is the rising trend. The third segment is another falling trend that has ending value of 103.2.

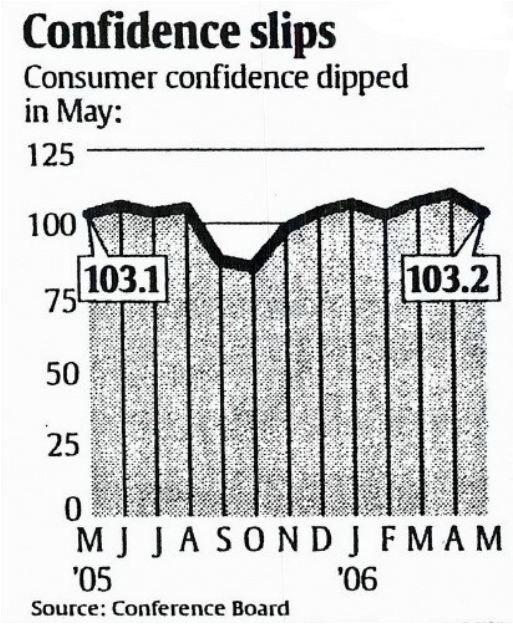


Figure 8-6: Example of a line graph used in the experiment (graph L26).

Table 8-8: Phase 2 Experiment results - test with visually impaired participants (Correct in %).

Graph code	# of questions answered	# of wrong or incomplete answers	% Correct answers
L3	4	0	100
L6	12	2	83.3
***L17**	**12**	**9**	**25**
L18	8	2	75
L21	8	1	87.5
***L26**	**8**	**5**	**37.5**
L28	12	0	100
L89	8	1	87.5
*L95	8	0	100

Table 8-9: Phase 3 Experiment results - control answers collected from sighted users from just glancing at the graphic (Correct in %).

Graph code	# of questions answered	# of answers that required going back to the graph	% Correct answers
L3	9	3	66.7
L6	23	4	62.5
*L17	27	14	25.9
L18	17	12	5.6
L21	19	7	50
*L26	15	6	18.8
L28	29	8	53.3
L89	20	11	35
*L95	19	15	20

Table 8-10: Phase 3 Experiment results - control answers collected from sighted users after carefully examining the graph (Correct in %).

Graph code	# of questions answered	# of wrong or incomplete answers after carefully examining the graphic	% Correct answers
L3	9	0	100
L6	23	4	82.6
*L17	27	6	77.7
L18	17	4	76.4
L21	19	2	89.4
*L26	15	6	60
L28	29	5	82.7
L89	20	2	90
*L95	19	0	100

Even though 3 out of 4 visually impaired users decided to listen to the summary more than once, they still had difficulties memorizing the answers to some of the questions. Line graph L26 (Figure 8-6) had two questions and the summary did not provide the answer for one of them. The questions for this graph were:

1. Question 1: **What is the difference in consumer confidence in May of 2005 and May of 2006?** Answer: 0.1
2. Question 2: **Which month had the lowest consumer confidence in 2005 and 2006?** Answer: October 2005

Line graph L28 showed an interesting result. For that graph, all of the answers provided by users with visual impairments were correct, while not all of the answers provided by sighted users for the same graph were correct. Figure 8-7shows the graph followed by its summary:

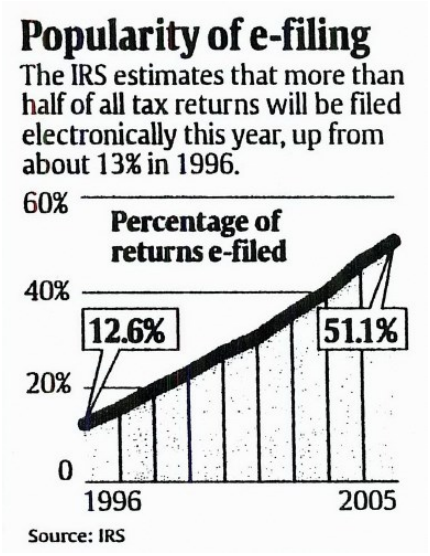


Figure 8-7: Example of a line graph used in the experiment (graph L28).

Summary generated for the graph in Figure 8-7:

The image shows a line graph, which presents IRS's percentage of returns e-filed. The line graph shows a rising trend from 1996 to 2005. The rising trend has a starting value of 12.6 percent and an ending value of 51.1 percent.

For this graph, one of the questions was: *By how much did the percentage of retiring e-files grow from 1996 to 2005?* Some of the sighted users decided not to go back before answering it and all of the users with visual impairments remembered the values and were able to answer the question correctly.

The average percent of correct answers from the visually impaired users was 75%, while the average percentage of correct answers from sighted users was 80.87%. From instances where sighted users got the answers wrong, we noticed that most of the time the questions asked for specific numbers for which they preferred to guess the answer instead of going back to view the graphic image again. For some questions

where blind users performed better than sighted, we could see from the sighted users' comments that they thought the questions were hard and/or the information was not easily accessible. For example, there were cases where the question asked for maximum or minimum values and those were not annotated, making an interpolation necessary, whereas the system provided these numbers in some of the summaries.

From the comments provided by the blind participants, we could see that individual preferences were stated (one participant declared that he would prefer the summary to have less information than what was provided, allowing the user to ask for more if he/she preferred). This consideration is aligned with our intention of providing follow-up responses as described in Chapter 9. Most of the comments provided by the blind participants stated that the summaries were clear and concise. The participants, in general, appreciated the clarity with which the information was delivered to them.

By analyzing the statistical significance of the data, we noticed that the null hypothesis is true when comparing right and wrong answers coming from blind and sighted participants. This result shows that the visual impairment was not a factor when the blind participants answered a question wrong or right when compared to sighted participants. Another possibility is that the amount of input data is not enough to draw conclusions, which would require that a large number of blind participants be recruited and a larger experiment be performed. Table 8-11 shows the statistical significance results where a p-value of 0.09908 is yielded.

Table 8-11: Statistical significance data for blind and sighted users' answers.

	Blind users	**Sighted users**	**Total**	**Prob**
Correct answers	60	149	209	0.81
Wrong answers	20	29	49	0.19
Total	80	178	258	

8.3 Thought Experiment

Another possible experiment might contain the same tasks, but compare the results of our system with those obtained from a baseline. Such baseline does not exist. One possibility could be to provide them with summaries generated using Benetech (Benetech, 2016) guidelines for line graph description as they are made available.

8.4 Summary

This chapter presented the evaluation of the generated summaries for line graphs at different reading levels and the usefulness of the summaries content for visually impaired users. It describes four different experiments performed in order to capture how well the system is able to adapt its generated text to a target reading level and how these different text complexities affect the understanding of it by readers at different reading levels.

The results for the automatic adaptation of the generated summaries to the desired reading levels was achieved successfully. In addition, experiments with human subjects were performed in order to assess if the perception human readers had

about different text complexities was in accordance with the decisions the system made to generate summaries at different grade levels. Results show that users agreed with the same order of text complexity generated by the system more than 70% of the time. In order to assess whether summaries generated at different reading levels would be preferred by users at that reading level, evaluations were also performed with 5th graders and college students. Conclusions were that not all users have the same reading preferences with regards to readability level, even when they are at the same school grade level. This confirmed our contention that the generated text should follow the reading level of the article in which the graph appears, since presumably the user is comfortable with the reading level used by the venue.

The final evaluation presented in this chapter was performed with visually impaired users. The goal was to assess how well the system could deliver useful information so that these users could perform specific tasks. In this specific case, the task was to verify how well they could answer questions related to the graphic by having access to the summaries when compared to how well sighted users were able to answer the same questions when looking at the graphic images. The results show that the system successfully provides the information needed by visually impaired users in order for them to be able to answer important questions about the graphic.

The system generates a summary that aims to deliver the most important information conveyed by the graphic. The summaries are adapted to the reading level of the articles in which the graphics appear, granting access to users at different reading levels. We envision that in the future SIGHT might be used for generating Alt-text for informational graphics.

Chapter 9

CONCLUSIONS AND FUTURE RESEARCH

This thesis presented the work done on the SIGHT system in order to provide sight-impaired users with access to line graphs which are available in popular media. Prior to this work, the SIGHT system provided generated textual summaries for simple bar charts found online. From the accessibility perspective, this work contributed by allowing visually impaired users to also, through the use of the system, have access to line graphs which are part of multimodal documents from newspapers and magazines.

Modifications were made and a functionality was added in order to ease the way users can install and use the system. Instead of having to install the whole system locally, a Chrome plugin was made available which allows the browser to submit a request to a server running the system. The image of the line graph is then sent to the server and the generated summary is sent back to the client once the request is completed on the server side.

The majority of the contributions of this work are to the Natural Language Generation field of research. The content determination component, for example, was expanded in order to consider additional information about the graph. For this version of the SIGHT system, secondary messages with non-trivial probabilities identified by the Intention Recognition Module are also considered when deciding on the content of the summaries. Additionally, by using a graph-based approach for content determination which considers the importance of the features being described, the system allows the summaries of different graphs to be customized based on its

particular characteristics: the more salient features a graphic has, the more detailed is the summary that describes it.

Another contribution is the text organization phase where the content selected during the previous phase defines the ordering in which the propositions will be presented in the summary. Once again, the presence of a candidate message with non-trivial probability, or the number of salient features present in a specific trend of the graphic, might lead the organization module to emphasize that piece of information in the summary by conveying the most important trend first.

The most interesting and challenging contribution regards the micro planning phase employed during this work. For this phase, two main contributions were made. The first is the aggregation module and the decision about aggregating text differently for different reading levels. For this aspect, the system learned the characteristics of text that make it complex using a decision tree with annotated corpora. Based on these learned characteristics, the system uses a graph search in order to find a goal node (a realized summary) which is close in reading level to the target reading level. The search is guided by a heuristic which reflects the measurements of the characteristics obtained in the learning phase.

The second contribution of the micro planning phase is the way the system decides on the lexical items that can be appropriate to the context of line graphs and also be suitable to the target reading level. Due to the lack of a readily available set of synonyms that are appropriate for describing line graphs and to the need for a set of synonyms from which the system can pick lexical items, a novel approach was developed. For lexicalization, an initial set of seed words is used in order to expand synonyms using a thesaurus. After this first step, the system filters the synonyms that

are appropriate for describing line graphs by considering co-occurrence in the 5-gram Google Books corpus and top word-vector similarity from word2vec.

The evaluations presented in the previous chapter show that the system successfully generates text at different reading levels, that these different texts are also perceived by human judges as belonging to different grade levels, and that human readers at different reading levels also prefer text that is generated at their respective reading levels. The system is also able to successfully deliver the most important message of line graphs to users who are visually impaired, allowing them to perform tasks sighted users did while accessing the graphic images.

We consider this work to be a valuable resource for research in the area of graph accessibility and Natural Language Generation. The next section addresses issues that can still be explored in both areas.

9.1 Future Work from the Natural Language Generation Perspective

9.1.1 Pronominalization

An important feature of coherent and understandable text is the pronominalization of referring expressions, which avoids reintroduction of entities every time they are mentioned. The experiment mentioned in Chapter 5 showed that the reintroduction of entities or the repetition of referring expressions (when a pronoun cannot be used) in fact jeopardized the understanding of some passages in the summaries. The participants would usually complain that a given summary was confusing because it could be "better presented" and they would additionally provide us with comments regarding the reintroduction of the referring expressions. From these results, we concluded that it would be valuable to include a pronominalization

phase after the aggregation phase so that even the summaries that are at a lower grade level would not repeat the referring expression when using multiple non-aggregated sentences.

The propositions chosen by the content determination framework contain the information about their "parents" (features such as volatility and steepness point to the trend of the graphic they belong to). This relationship is the clue used to define discourse focus. Such information can be used by a pronominalization module, when implemented.

9.1.2 Coordinated lexicalization in a summary

Two aspects can be explored with regards to the coordination of lexical items within a summary. The first is the ability to choose different lexical items to describe the same concept in the summary by using a different referring expression when, for example, reintroducing an entity. One example is that **trend** and **segment** can both be used to describe the concept **trend**, but caution is needed in order to assure that the reader understands that they are both referring to the same entity.

The second aspect is the coordination of contrasting concepts. The results of the experiment performed with 5th graders and college students showed that participants would prefer to see top vs bottom, maximum vs minimum, first vs last, higher vs lower, instead of seeing randomly selected lexical items being used to realize opposite concepts. This is an interesting topic for future work on the lexicalization phase.

9.2 Future Work from the Accessibility Perspective

9.2.1 Grouped Bar Charts

Another important aspect in the extension of the SIGHT system is to handle other types of graphics. Multiple line graphs, grouped bar charts, and pie charts, are all graphs that possess a different set of intended messages and are used for different purposes, opening the problem of how a system should describe them.

Grouped bar charts, specifically, have been the object of study for recognizing their intended messages presented in (Burns et al., 2010; Burns, Carberry, & Schwartz, 2013a, 2013b). Exploring this type of graph might unveil a completely different set of challenges from the Natural Language Generation perspective. Specifically, it is anticipated that these graphs pose challenges for organizing a summary due to their complexity. Therefore, enabling the SIGHT system to also generate summaries for grouped bar charts will actually contribute to both fields of research (accessibility and NLG).

9.2.2 Follow up responses

Our plans for future work on the SIGHT system includes the extension of the generation module to allow the user to access follow-up information about the graphic. Since the initial summary only delivers the most important information (intended message + secondary messages + outstanding visual features), follow-up responses should be able to deliver relevant in-depth information based upon a request from the user. Challenges here involve the continuous nature of line graphs and methods for identifying what additional information to provide.

REFERENCES

Alty, J. L., & Rigas, D. (2005). Exploring the Use of Structured Musical Stimuli to Communicate Simple Diagrams: The Role of Context. *International Journal of Human-Computer Studies, 62*(1), 21-40.

Barzilay, R., & Lapata, M. (2006). *Aggregation Via Set Partitioning for Natural Language Generation.* Paper presented at the Human Language Technologies - North American Chapter of the Association for Computational Linguistics (HLT-NAACL) New York City, NY.

Bayyarapu, H. S. (2011). *Efficient algorithm for Context Sensitive Aggregation in Natural Language generation.* Paper presented at the Recent Advances in Natural Language Processing (RANLP), Hissar, Bulgaria.

Benetech. (2016). Benetech.

Bouayad-Agha, N., Casamayor, G., & Wanner, L. (2011). *Content selection from an ontology-based knowledge base for the generation of football summaries.* Paper presented at the Proceedings of the 13th European Workshop on Natural Language Generation, Stroudsburg, PA, USA.

Bouayad-Agha, N., Casamayor, G., Wanner, L., Fernando, D., & López, H. S. (2011). *FootbOWL: using a generic ontology of football competition for planning match summaries.* Paper presented at the Proceedings of the 8th Extended Semantic Web Conference on the Semantic Web: Research and Applications - Volume Part I, Berlin, Heidelberg.

Boyd, S. (1998). *TREND: A System for Generating Intelligent Descriptions of Time-Series Data.* Paper presented at the IEEE International Conference on Intelligent Processing Systems (ICIPS-1998).

Brown, L. M., & Brewster, S. A. (2003). *Drawing by Ear: Interpreting Sonified Line Graphs.* Paper presented at the International Conference on Auditory Display (ICAD), Boston, MA.

Burns, R., Carberry, S., & Elzer, S. (2010). *Visual and spatial factors in a bayesian reasoning framework for the recognition of intended messages in grouped bar charts.* Paper presented at the Proceedings of the AAAI Workshop on Visual Representations and Reasoning.

Burns, R., Carberry, S., & Schwartz, S. E. (2013a). *Modeling a Graph Viewer's Effort in Recognizing Messages Conveyed by Grouped Bar Charts.* Paper presented at the Conference on User Modeling, Adaptation and Personalization (UMAP), Rome, Italy.

Carberry, S., Elzer, S., & Demir, S. (2006). *Information graphics: an untapped resource for digital libraries.* Paper presented at the Proceedings of the 29th

annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA.
Carberry, S., Elzer Schwartz, S., McCoy, K., Demir, S., Wu, P., Greenbacker, C., . . . Moraes, P. (2013). Access to multimodal articles for individuals with sight impairments. *ACM Trans. Interact. Intell. Syst., 2*(4), 21:21-21:49.
Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., & Tait, J. (1999). *Simplifying Text for Language-Impaired Readers.* Paper presented at the Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL).
Center, P. R. (2010a). News Use Across Social Media Platforms.
Center, P. R. (2010b). Pew Internet and American Life Project: World Wide Web.
Chester, D., & Elzer, S. (2005). *Getting computers to see information graphics so users do not have to.* Paper presented at the Proceedings of the 15th International Symposium on Methodologies for Intelligent Systems.
Cohen, R. F., Haven, V., Lanzoni, J. A., Meacham, A., Skaff, J., & Wissell, M. (2006). *Using an audio interface to assist users who are visually impaired with steering tasks.* Paper presented at the Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility.
Cohen, R. F., & Yu, R. (2005). *PLUMB: Displaying Graphs to the Blind Using an Active Auditory Interface.* Paper presented at the Proceedings of the 7th international ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'05).
Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology, 60*, 283-284.
Collins-Thompson, K., & Callan, J. (2005). Predicting Reading Difficulty with Statistical Language Models. *J. Am. Soc. Inf. Sci. Technol., 56*(13), 1448-1462.
Collins-Thompson, K., & Callan, J. P. (2004). *A Language Modeling Approach to Predicting Reading Difficulty.* Paper presented at the Human Language Technologies - North American Chapter of the Association for Computational Linguistics (HLT-NAACL).
Common Core State Standards Initiative. (2014). Retrieved from http://www.corestandards.org/
Covington, M. A., He, C., Brown, C., Naçi, L., Brown, J., level Scale, A. D., & Plc, G. (2006). How Complex Is That Sentence? A Proposed Revision of the Rosenberg and Abbeduto D-level scale. Retrieved from the Computer Analysis of Speech for Psychological Research Report 2006-01.
Demir, S. (2010). *Sight for visually impaired users: Summarizing information graphics textually.* (Unpublished doctoral dissertation). University of Delaware, Delaware USA.
Demir, S., Carberry, S., & Elzer, S. (2007). *Effectively Realizing the Inferred Message of an Information Graphic.* Paper presented at the Proceedings of Recent Advances in Natural Language Processing Conference.

Demir, S., Carberry, S., & McCoy, K. F. (2008). *Generating textual summaries of bar charts.* Paper presented at the Proceedings of the Fifth International Natural Language Generation Conference, Stroudsburg, PA, USA.

Demir, S., Carberry, S., & McCoy, K. F. (2010). *A discourse-aware graph-based content-selection framework.* Paper presented at the Proceedings of the 6th International Natural Language Generation Conference, Stroudsburg, PA, USA.

Demir, S., Carberry, S., & McCoy, K. F. (2012). Summarizing Information Graphics Textually. *Computational Linguistics, 38*(3), 527-574.

Demir, S., Oliver, D., Schwartz, E., Elzer, S., Carberry, S., & McCoy, K. F. (2010). *Interactive SIGHT into information graphics.* Paper presented at the Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A), New York, NY, USA.

Demir, S., Oliver, D., Schwartz, E., Elzer, S., Carberry, S., McCoy, K. F., & Chester, D. (2010). Interactive SIGHT: textual access to simple bar charts. *New Rev. Hypermedia Multimedia, 16*, 245-279.

Demsar, J., Curk, T. v., Erjavec, A. v., Gorup, v. r., Hoč, e. T. v., Milutinovič, M., . . . Zupan, B. v. (2013). Orange: Data Mining Toolbox in Python. *J. Mach. Learn. Res., 14*(1), 2349-2353.

Dictionary.com, L. L. C. (2015). www.thesaurus.com.

Duboue, P. A., & McKeown, K. R. (2003). *Statistical acquisition of content selection rules for natural language generation.* Paper presented at the Proceedings of the 2003 conference on Empirical methods in natural language processing, Stroudsburg, PA, USA.

Elhadad, M., & Robin, J. (1998). *SURGE: a Comprehensive Plug-in Syntactic Realization Component for Text Generation*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.34.4728

Elzer, S., Carberry, S., Zukerman, I., Chester, D., Green, N., & Demir, S. (2005). *A Probabilistic Framework for Recognizing Intention in Information Graphics.* Paper presented at the Proceedings of the International Joint Conference on Artificial Intelligence.

Elzer, S., Green, N., Carberry, S., Carberry, R., & McCoy, K. (2003). *Extending Plan Inference Techniques to Recognize Intentions In Information.* Paper presented at the Proceedings of the Ninth International Conference on User Modeling.

Elzer, S., Schwartz, E., Carberry, S., Chester, D., Demir, S., & Wu, P. (2007). *A Browser Extension For Providing Visually Impaired Users Access To The Content Of Bar Charts On The Web.* Paper presented at the Proceedings of the International Conference on Web Information Systems and Technologies.

Elzer, S., Schwartz, E., Carberry, S., Chester, D., Demir, S., & Wu, P. (2008). *Accessible Bar Charts for Visually Impaired Users.* Paper presented at the Proceedings of the International Association of Science and Technology for Development (IASTED) International Conference on Telehealth/AT.

Fellbaum, C. (1998). *WordNet: An electronic Lexical Database*: The MIT Press.

Ferres, L., Lindgaard, G., Sumegi, L., & Tsuji, B. (2013). Evaluating a Tool for Improving Accessibility to Charts and Graphs. *ACM Trans. Comput.-Hum. Interact., 20*(5), 28:21-28:32.

Ferres, L., Parush, A., Roberts, S., & Lindgaard, G. (2006). *Helping people with visual impairments gain access to graphical information through natural language: The iGraph system.* Paper presented at the Proceeding of the 10th International Conference on Computers Helping People with Special Needs. Lecture Notes in Computer Science.

Ferres, L., Verkhogliad, P., Lindgaard, G., Boucher, L., Chretien, A., & Lachance, M. (2007a). *Improving accessibility to statistical graphs: the iGraph-Lite system.* Paper presented at the Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility, New York, NY, USA.

Ferres, L., Verkhogliad, P., Lindgaard, G., Boucher, L., Chretien, A., & Lachance, M. (2007b). *Improving Accessibility to Statistical Graphs: the inspectGraph System.* Paper presented at the Proceedings of the Ninth International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS).

François, T., & Miltsakaki, E. (2012). *Do NLP and Machine Learning Improve Traditional Readability Formulas?* Paper presented at the Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations, Stroudsburg, PA, USA.

Fredj, Z. B., & Duce, D. A. (2007). GraSSML: accessible smart schematic diagrams for all. *Universal Access in the Information Society, 6*(3), 233-247.

Fsf. (2005). Style and Diction GNU project.

Gatt, A., Portet, F., Reiter, E., Hunter, J., Mahamood, S., Moncur, W., & Sripada, S. (2009). From data to text in the Neonatal Intensive Care Unit: Using NLG technology for decision support and information management. *AI Communications, 22*(3), 153-186.

Gatt, A., & Reiter, E. (2009). *SimpleNLG: A Realisation Engine for Practical Applications.* Paper presented at the Proceedings of the 12th European Workshop on Natural Language Generation, Stroudsburg, PA, USA.

Goncu, C., & Marriott, K. (2008). *Tactile chart generation tool.* Paper presented at the Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility, New York, NY, USA.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., Cai, Z., Dempsey, K., Floyd, Y., . . . Correspondence, F. Y. (2004). *Coh-Metrix: Analysis of text on cohesion and language.* Paper presented at the M. Louwerse Topics in Cognitive Science.

Greenbacker, C., Carberry, S., & McCoy, K. (2011, July). *A Corpus of Human-written Summaries of Line Graphs.* Paper presented at the Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop, Edinburgh, Scotland.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). *Combining Lexical and Grammatical Features to Improve Readability Measures for First*

and Second Language Texts. Paper presented at the Human Language Technologies - North American Chapter of the Association for Computational Linguistics (HLT-NAACL).

Heilman, M., Collins-Thompson, K., & Eskenazi, M. (2008). *An Analysis of Statistical Models and Features for Reading Difficulty Prediction.* Paper presented at the Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, Stroudsburg, PA, USA.

Hovy, E. H. (1988). *Planning coherent multisentential text.* Paper presented at the Proceedings of the 26th annual meeting on Association for Computational Linguistics, Stroudsburg, PA, USA.

Jordan, P. W., & Walker, M. A. (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research, 24*, 157-194.

Kanungo, T., & Orr, D. (2009). *Predicting the Readability of Short Web Summaries.* Paper presented at the Proceedings of the Second ACM International Conference on Web Search and Data Mining, New York, NY, USA.

Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., . . . Welty, C. (2010). *Learning to Predict Readability using Diverse Linguistic Features.* Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010.

Kennel, A. R. (1996). *Audiograf: a diagram-reader for the blind.* Paper presented at the Proceedings of the second annual ACM conference on Assistive technologies, New York, NY, USA.

Kincaid, J. P., Fishburne, R. P., 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*. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED108134

Kipper, K., Dang, H. T., & Palmer, M. (2000). *Class-Based Construction of a Verb Lexicon.* Paper presented at the Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence.

Klapaftis, I. P., & Manandhar, S. (2005). *Google and wordnet based word sense disambiguation.* Paper presented at the Proceedings of the 22nd International Conference on Machine Learning Workshop on Learning and Data Mining.

Koopmans, L. H., Owen, D. B., & Rosenblatt, J. I. (1964). Confidence intervals for the coefficient of variation for the normal and log normal distributions. *Biometrika, 51*(1-2), 25-32.

Krufka, S. E., & Barner, K. E. (2006). A user study on tactile graphic generation methods. *Behaviour & Information Technology, 25*(4), 297-311.

Kurze, M. (1996). *TDraw: a computer-based tactile drawing tool for blind people.* Paper presented at the Proceedings of the second annual ACM conference on Assistive technologies.

Laughlin, G. H. M. (1969). SMOG Grading-a New Readability Formula. *Journal of Reading, 12*(8), pp. 639-646.
Lexile Framework for Reading. (2015).
Library, A. (2015, August). Austin Public Library Electronic Catalog. Retrieved from http://library.austintexas.gov/
Loper, E., & Bird, S. (2002). *NLTK: The Natural Language Toolkit.* Paper presented at the Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1, Stroudsburg, PA, USA.
Louis, A., Joshi, A., & Nenkova, A. (2010). *Discourse indicators for content selection in summarization.* Paper presented at the Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Stroudsburg, PA, USA.
Mann, W., & Thompson, S. (1987). Rhetorical Structure Theory: Description and Construction of Text Structures. In G. Kempen (Ed.), *Natural Language Generation* (Vol. 135, pp. 85-95): Springer Netherlands.
Mann, W. C., & Thompson, S. A. (1987). Rhetorical structure theory: A theory of text organization. In L. Polanyi (Ed.), *The Structure of Discourse*. Norwood, N.J.: Ablex Publishing Corporation.
McGookin, D. K., & Brewster, S. A. (2006). *SoundBar: exploiting multiple views in multimodal graph browsing.* Paper presented at the Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles, New York, NY, USA.
McKeown, K. (1992). *Text Generation*: Cambridge University Press.
McKeown, K. R. (1985). Discourse strategies for generating natural-language text. *Artificial Intelligence, 27*(1), 1-41.
Merriam, W. (2016). Merriam-Webster.
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.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Computing Research Repository* (*CoRR), abs/1301.3781*.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality.* Paper presented at the Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.
Moore, J. D., & Paris, C. e. c. L. (1993). Planning text for advisory dialogues: capturing intentional and rhetorical information. *Comput. Linguist., 19*(4), 651-694.

Moraes, P., McCoy, K., & Carberry, S. (2014b). *Adapting Graph Summaries to the Users' Reading Levels.* Paper presented at the Proceedings of the 8th International Natural Language Generation Conference.
Moraes, P., Sina, G., McCoy, K., & Carberry, S. (2014). *Generating Summaries of Line Graphs.* Paper presented at the Proceedings of the 8th International Natural Language Generation Conference.
Napolitano, D., Sheehan, K., & Mundkowsky, R. (2015, June). *Online Readability and Text Complexity Analysis with TextEvaluator.* Paper presented at the Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, Denver, Colorado.
NFB. (2013). National Federation of the Blind.
Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*. Retrieved from http://ilpubs.stanford.edu:8090/422/
Pitler, E., & Nenkova, A. (2008). *Revisiting Readability: A Unified Framework for Predicting Text Quality.* Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA.
Ramloll, R., Yu, W., Brewster, S. A., Riedel, B., Burton, A. M., & Dimigen, G. (2000). *Constructive sonified haptic line graphs for the blind student: First steps.* Paper presented at the Fourth Annual ACM Conference on Assistive Technologies, Arlington, VA.
Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Nat. Lang. Eng., 3*(1), 57-87.
Reiter, E., & Dale, R. (2000). Building Natural-Language Generation Systems. *Cambridge University Press*.
Reiter, E., Sripada, S., & Robertson, R. (2003). Acquiring Correct Knowledge for Natural Language Generation. *Journal of Artificial Intelligence Research, 18*, 491-516.
Rello, L., & Baeza-Yates, R. (2014). *Evaluation of DysWebxia: A Reading App Designed for People with Dyslexia.* Paper presented at the Proceedings of the 11th Web for All Conference, New York, NY, USA.
Rello, L., Baeza-Yates, R., Bott, S., & Saggion, H. (2013). *Simplify or Help?: Text Simplification Strategies for People with Dyslexia.* Paper presented at the Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, New York, NY, USA.
Rello, L., & Baeza-Yates, R. A. (2012). The presence of English and Spanish dyslexia in the Web. *The New Review of Hypermedia and Multimedia, 18*(3), 131-158.
Russell, S. J., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2 ed.): Pearson Education.

Saggion, H., vS, t. S., Bott, S., Mille, S., Rello, L., & Drndarevic, B. (2015). Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Trans. Access. Comput., 6*(4), 14:11-14:36.

Schwarm, S. E., & Ostendorf, M. (2005). *Reading Level Assessment Using Support Vector Machines and Statistical Language Models.* Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA.

Sheehan, K. M., Kostin, I., Futagi, Y., & Flor, M. (2010). *Generating automated text complexity classifications that are aligned with targeted text complexity standards*. Retrieved from https://www.ets.org/Media/Research/pdf/RR-10-28.pdf.

Si, L., & Callan, J. (2001). *A Statistical Model for Scientific Readability.* Paper presented at the Proceedings of the Tenth International Conference on Information and Knowledge Management, New York, NY, USA.

Siddharthan, A. (2003). *Preserving Discourse Structure when Simplifying Text.* Paper presented at the Proceedings of the 2003 European Natural Language Generation Workshop.

Sinha, R., & Mihalcea, R. (2007). *Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity.* Paper presented at the Proceedings of the International Conference on Semantic Computing, Washington, DC, USA.

Smith, E. A., & Senter, R. J. (1967). Automated Readability Index.

Stajner, S., Mitkov, R., & Saggion, H. (2014, April). *One Step Closer to Automatic Evaluation of Text Simplification Systems.* Paper presented at the Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), Gothenburg, Sweden.

Tanaka-Ishii, K., Tezuka, S., & Terada, H. (2010). Sorting Texts by Readability. *Comput. Linguist., 36*(2), 203-227.

Temnikova, I., & Maneva, G. (2013, August). *The C-Score -- Proposing a Reading Comprehension Metrics as a Common Evaluation Measure for Text Simplification.* Paper presented at the Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations, Sofia, Bulgaria.

Vajjala, S., & Meurers, D. (2012). *On Improving the Accuracy of Readability Classification Using Insights from Second Language Acquisition.* Paper presented at the Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, Stroudsburg, PA, USA.

Walker, M. A., Rambow, O., & Rogati, M. (2001). *SPoT: a trainable sentence planner.* Paper presented at the Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, Stroudsburg, PA, USA.

Wall, S., & Brewster, S. (2006). *Feeling What You Hear: Tactile Feedback for Navigation of Audio Graphs.* Paper presented at the Proceedings of the

SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA.
Wikipedia. (2004). Wikipedia, The Free Encyclopedia.
Wilkinson, J. (1995). *Aggregation in Natural Language Generation: Another Look*. Retrieved from
Williams, S., & Reiter, E. (2004, August). *Reading errors made by skilled and unskilled readers: evaluating a system that generates reports for people with poor literacy.* Paper presented at the Fourteenth Annual Meeting of the Society for Text and Discourse, Chicago.
Williams, S., & Reiter, E. (2005a). *Appropriate Microplanning Choices for Low-Skilled Readers.* Paper presented at the IJCAI.
Williams, S., & Reiter, E. (2005b). *Generating readable texts for readers with low basic skills.* Paper presented at the Proceedings of the 10th European Workshop on Natural Language Generation (EWNLG 2005).
Williams, S., & Reiter, E. (2008). Generating basic skills reports for low-skilled readers. *Natural Language Engineering, 14*(4), 495-525.
Williams, S., Reiter, E., & Osman, L. (2003, August). *Experiments with discourse-level choices and readability.* Paper presented at the Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003), Budapest.
Wu, P., Carberry, S., & Elzer, S. (2010). *Segmenting Line Graphs into Trends.* Paper presented at the Proceedings of the Twelth International Conference on Artificial Intelligence.
Wu, P., Carberry, S., Elzer, S., & Chester, D. (2010). *Recognizing the intended message of line graphs.* Paper presented at the Proceedings of the 6th international conference on Diagrammatic representation and inference, Berlin, Heidelberg.
Xing, W., & Ghorbani, A. (2004). *Weighted PageRank algorithm.* Paper presented at the Communication Networks and Services Research, May, 2004.
Yu, W., & Brewster, S. (2002). *Comparing two haptic interfaces for multimodal graph rendering.* Paper presented at the Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2002. HAPTICS 2002. Proceedings. 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems.
Yu, W., Kangas, K., & Brewster, S. (2003). *Web-based haptic applications for blind people to create virtual graphs.* Paper presented at the Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2003. HAPTICS 2003.
Yu, W., Ramloll, R., & Brewster, S. A. (2001). *Haptic Graphs for Blind Computer Users.* Paper presented at the Proceedings of the First International Workshop on Haptic Human-Computer Interaction, London, UK.
Zhou, M. X., & Feiner, S. K. (1997). *Top-down hierarchical planning of coherent visual discourse.* Paper presented at the Proceedings of the 2nd international conference on Intelligent user interfaces, New York, NY, USA.

Appendix A

PROPOSITION REALIZATION TEMPLATES

Each of the proposition templates below can be used to realize one or more propositions from the set listed in Section 7.4. They show the semantic representation of the different templates using the notation **root_predicate(set_of_arguments)**. The set of propositions which use the template is listed next. The possible realizations a template is allowed is also listed. It uses the notation <arg1> show-predicate <arg2> where <x> means realization of x; x_y means the already-existing realization of x and it is of syntactic category y; [[x]] means that x is an optional argument. The base lexicon of the root predicate used to perform synonym expansion – described in Section 7.7.

Some propositions such as **graph_volatility, trend_volatility** and **trend_steepness** are complex as they show to have more constraints with the types of lexicon used to describe these phenomena of the graph. This also affects the number of possible ways these propositions can be realized.

Show template

show(arg1, arg2)

type arg1: *entity*

type arg2: *entity*

Set of propositions which use this template:

graph_type, graph_overall_behaviour, composed_trend

Possible realizations:

- As a sentence:

  <arg1> show-predicate <arg2>

  <arg1> is an NP

  <arg2> is an NP

- As a subordinating conjunction:

  arg1$_{NP}$ which shows <arg2>

  <arg2> is an NP

Base lexicon:

Show-predicate: *verb(show)*

Type template

type(arg1, arg2).

type arg1: *entity*

type arg2: *entity*

Set of propositions which use this template:

trend_description

Possible realizations:

- As a sentence:

  <arg1> type-predicate <arg2>

  <arg1> is an NP

  <arg2> is an NP

- As a subordinating conjunction:

  arg1$_{NP}$ which is <arg2>

  <arg2> is an NP

Base lexicon:

Type-predicate: *verb(is)*

Present template

present(arg1, arg2, [[arg3]])

type arg1: *entity*

type arg2: *entity*

type arg3 (optional): *unit and scale*

Set of propositions which use this template:

entity_description

Possible realizations:

- As a sentence:

  <arg1> present-predicate <arg2>

  <arg1> is an NP

  <arg2> is an NP

  <arg3> is a PP

- As a subordinating conjunction:

  arg1$_{NP}$ which presents <arg2>[[<arg3>]]

  <arg2> is an NP

  <arg3> is a PP

Base lexicon:

Present-predicate: *verb(present)*

Volatile template

volatile(arg1, [[arg2]])

type arg1: *entity*

type arg2 (optional): *volatility degree*

Set of propositions which use this template:

graph_volatility, trend_volatility

Possible realizations:

- As a sentence:

  <arg1> is [[arg2]] volatile-word

  <arg1> is an NP

  <arg2> is degree_ADV_ADJ

  volatile-word is volatile_adjective

  <arg1> shows [[arg2]] volatile-word

  <arg1> is an NP

  <arg2> is degree_ADV_Noun

  volatile-word is volatile_noun

- As an adjective:

  [[arg2]] volatile {arg1}NP

  <arg2> is degree_ADV_ADJ

  volatile-word is volatile_adjective

- As a subordinating conjunction:

  {arg1}NP which is | show [[arg2]] volatile-word

  {arg1}NP which is [arg2] volatile-word

  <arg2> is degree_ADV_ADJ

  volatile-word is volatile_adjective

{arg1}NP which shows [[arg2]] volatile-word

<arg2> is degree_ADV_Noun

volatile-word is volatile_noun

Base lexicon:

Volatile-word: *adjective(jagged)*

Steep template

steep(arg1, [[arg2]]) – *The rising trend is steep.*

type arg1: *entity*

type arg2 (optional): *steepness degree*

Set of propositions which use this template:

trend_steepness

Possible realizations:

- As a sentence:

  <arg1> is [[arg2]] steep-word

  <arg1> is an NP

  <arg2> is degree_ADV_ADJ

  steep -word is steep _adjective

  <arg1> shows [[arg2]] steep -word

  <arg1> is an NP

  <arg2> is degree_ADV_Noun

  steep -word is steep _noun

- As an adjective:

  [[arg2]] steep {arg1}NP

<arg2> is degree_ADV_ADJ

steep -word is steep _adjective

- As a subordinating conjunction:

{arg1}NP which is | show [[arg2]] steep -word

{arg1}NP which is [arg2] steep -word

<arg2> is degree_ADV_ADJ

steep -word is steep _adjective

{arg1}NP which shows [[arg2]] steep -word

<arg2> is degree_ADV_Noun

steep -word is steep _noun

Base lexicon:

Steep-word: *adjective(steep)*

Value template

value(arg1, arg2, [[arg3]])

type arg1: *entity*

type arg2: *entity*

type arg3 (optional): *unit and scale*

Set of propositions which use this template:

graph_initial_value, graph_end_value, trend_initial_value, trend_end_value

Possible realizations:

- As a sentence:

<arg1> have-predicate <arg2>[[arg3]]

<arg1> is an NP

<arg2> is an NP

<arg3> is a PP

- As a subordinating conjunction:

arg1$_{NP}$ which has <arg2>[[<arg3>]]

<arg2> is an NP

<arg3> is a PP

Base lexicon:

Have-predicate: *verb(have)*

Date template

date(arg1, arg2, arg3)

Set of propositions which use this template:

graph_initial_date, graph_end_date, trend_initial_date, trend_end_date

type arg1: *entity*

type arg2: *entity*

type arg3: *entity*

Possible realizations:

- As a sentence:

<arg1> have-predicate <arg2><arg3>

<arg1> is an NP

<arg2> is an NP

<arg3> is a PP

- As a subordinating conjunction:

arg1$_{NP}$ which has <arg2><arg3>

<arg2> is an NP

<arg3> is a PP

Base lexicon:

Have-predicate: *verb(have)*

Value_change template

value_change(arg1, arg2, arg3, [[arg4]])

type arg1: *entity*

type arg2: *entity*

type arg3: *entity*

type arg4 (optional): *unit and scale*

Set of propositions which use this template:

graph_absolute_change, graph_rate_change, trend_absolute_change, trend_rate_change

Possible realizations:

- As a sentence:

  <arg1> show-predicate <arg2><arg3>[[arg4]]

  <arg1> is an NP

  <arg2> is an NP

  <arg3> is a NP

  <arg4> is a PP

- As a subordinating conjunction:

  arg1$_{NP}$ which shows <arg2><arg3>[[<arg4>]]

<arg2> is an NP

<arg3> is a NP

<arg4> is a PP

Base lexicon:

Show-predicate: *verb(show)*

Date_change template

date_change(arg1, arg2, arg3)

Set of propositions which use this template:

graph_overall_period_years, graph_overall_period_months, graph_overall_period_days, trend_overall_period_years, trend_overall_period_months, trend_overall_period_days

Possible realizations:

- As a sentence:

  <arg1> span-predicate <arg2><arg3>

  <arg1> is an NP

  <arg2> is an NP

  <arg3> is a NP

- As a subordinating conjunction:

  arg1$_{NP}$ which spans <arg2><arg3>

  <arg2> is an NP

  <arg3> is a NP

Base lexicon:

Span-predicate: *verb(span)*

Max_min_value template

max_min_value(arg1, arg2, arg3,[[arg4]])

Set of propositions which use this template:

maximum_point_value, minimum_point_value

Possible realizations:

- As a sentence:

  <arg1> have-predicate <arg2><arg3>[[arg4]]

  <arg1> is an NP

  <arg2> is an NP

  <arg3> is a PP

  <arg4> is an NP

- As a subordinating conjunction:

  arg1$_{NP}$ which has <arg2><arg3>[[arg4]]

  <arg2> is an NP

  <arg3> is a PP

  <arg4> is an NP

Base lexicon:

Have-predicate: *verb(have)*

Max_min_date template

max_min_date(arg1, arg2) – *The maximum value occurs in 1982.*

Set of propositions which use this template:

maximum_point_date, minimum_point_date

Possible realizations:

- As a sentence:

  <arg1> occur-predicate <arg2>

  <arg1> is an NP

  <arg2> is an PP

- As a subordinating conjunction:

  arg1$_{NP}$ which occur <arg2>

  <arg2> is an PP

Base lexicon:

Occur-predicate: *verb(occur)*

Appendix B

READING LEVEL BASED LEXICON

The following table shows the reading level based lexicon created by the steps described in Section 7.7. For the generation of summaries, the lexical item used to describe a concept is randomly chosen before the graph search algorithm starts searching for the best aggregation plan. The reader will notice that the lexicon for the higher grade levels include the items from the lower grade levels. This is due to the assumption that if a lexical item is introduced at a lower grade level, it is appropriate to use it at a higher grade level.

Grade level based lexicon	
	Consist
4th – 5th	consist of contain include
6th – 8th	consist of contain include
9th – 10th	involve contain is composed of consist of

Grade level based lexicon	
	include
11th - college	involve contain is composed of consist of include
Decrease	
4th – 5th	reduce
6th – 8th	reduce
9th – 10th	reduce decline
11th - college	lower reduce decline
Falling	
4th – 5th	decreasing
6th – 8th	decreasing
9th – 10th	decreasing
11th - college	decreasing lowering
Final	
4th – 5th	last
6th – 8th	last

Grade level based lexicon	
9th – 10th	final last
11th - college	final last finished
Graph	
4th – 5th	graph
6th – 8th	graph diagram
9th – 10th	graph diagram
11th - college	graph diagram
Image	
4th – 5th	image
6th – 8th	image
9th – 10th	picture image drawing
11th - college	drawing reflection image picture

Grade level based lexicon	
Increase	
4th – 5th	increase
6th – 8th	extend increase
9th – 10th	extend rise increase
11th - college	extend rise increase
Initial	
4th – 5th	first early
6th – 8th	first early
9th – 10th	primary original first early
11th - college	initial first early primary

Grade level based lexicon	
	original
Jagged	
4th – 5th	variable
6th – 8th	variable
9th – 10th	variable jagged
11th - college	variable jagged
Maximum	
4th – 5th	top
6th – 8th	highest top
9th – 10th	highest top
11th - college	maximal maximum top highest
Minimum	
4th – 5th	minimum
6th – 8th	minimum
9th – 10th	least lowest

Grade level based lexicon	
	minimum
11th - college	least lowest minimum minimal
Occur	
4th – 5th	is found
6th – 8th	come about is found occur
9th – 10th	come about take place is found occur
11th - college	come about is present arise take place is found occur
Present	
4th – 5th	give
6th – 8th	give

Grade level based lexicon	
	present
9th – 10th	give present
11th - college	give present
Rising	
4th – 5th	rising
6th – 8th	rising
9th – 10th	rising ascending
11th - college	growing ascending increasing rising
Show	
4th – 5th	show
6th – 8th	show reveal
9th – 10th	demonstrate show reveal
11th - college	demonstrate show

Grade level based lexicon	
	reveal
Span	
4th – 5th	span
6th – 8th	span cover
9th – 10th	span cover
11th - college	span extend cover
Stable	
4th – 5th	solid
6th – 8th	solid constant
9th – 10th	solid steady constant
11th - college	solid steady constant
Steep	
4th – 5th	sharp high

Grade level based lexicon	
6th – 8th	sharp high
9th – 10th	sharp high
11th - college	Sharp precipitous steep high
Trend	
4th – 5th	direction
6th – 8th	trend direction
9th – 10th	trend direction orientation
11th - college	trend direction orientation drift

Appendix C

HUMAN INTELLIGENT TASK

Instructions

This HIT contains three parts:

1) Fragment evaluation warm-up: Read two fragments of text and choose which one you consider has the highest grade level from the two.

2) Read summaries and rank them from easiest to hardest by associating them with different grade levels.

3) Provide requested explanation about some of your answers.

Attention: FOR YOUR HIT TO BE ACCEPTED, PLEASE MAKE SURE THAT YOU DO THE FOLLOWING: WHILE ASSOCIATING SUMMARIES TO READING LEVELS, MAKE SURE YOU USE EACH GRADE LEVEL ONLY ONCE (DO NOT REPEAT A GRADE LEVEL FOR DIFFERENT SUMMARIES); ADDITIONALLY, MAKE SURE YOU PROVIDE A MEANINGFUL EXPLANATION FOR YOUR ANSWERS AT THE END OF THE TASK (ITEMS 6 AND 7).

Here are some additional instructions:

- Read the summaries and rank them from easiest to hardest by associating it to the appropriate grade level. The judgement of grade level should be made based on the comparison across summaries. For example, the summary that you perceive to be the easiest one should be associated with the 4th grade level.
- Because you are ranking the summaries, each grade level should be used only once. This means that no grade level can be repeated across answers.
- You will always be given 4 (four) summaries and 4 (four) grade levels.
- The summaries are NOT ordered in any way.
- If you believe any two summaries are really similar in terms of their ranking, you must still assign them two different grade levels.

1. Given the following two fragments of text, please choose which fragment you consider to have a higher grade level when compared to the other:

Fragment 1: Eyes turn light into sight with help from the brain and they work in the same way for people. By looking at this page you may think you see words and pictures but, believe it or not, you don't; all you see is light bouncing off the page and that is possible by the secret in the rules of light.

Fragment 2: With help from the brain, eyes turn light into sight. Eyes work in the same way for people. Look at this page. You may think you see words and pictures. Believe it or not, you don't. All you see is light bouncing off the page. How is this possible? The secret is in the rules of light.

Select the fragment that you believe has the highest grade level:

(Dropbox with options: Fragment 1 and Fragment 2)

2. First summary

A line graph which presents the number of annual difference from Seattle's 1899 sea levels in inches and consists of a changing trend composed of a stable trend from 1900 to 1928 followed by a rising trend through 2003 and shows much variability is shown by the image. A steady drift which has an initial value of 1.97 inches is given by the first segment. An increasing drift which has a final value of 8.9 inches is given by the second segment. The maximal value of 11 inches is reached by the graph which has the minimal value of 0.03 inches.

Select the grade level that best fits the summary above, in your opinion (remember that you can only use each grade level once):
(Dropbox with options: 4^{th} grade, 7^{th} grade, 10^{th} grade, College level)

3. Second summary
There is an image. The image shows a line graph. The graph gives the number of annual difference from Seattle's 1899 sea levels in inches. The graph consists of a changing trend composed of a stable trend from 1900 to 1928 followed by a rising trend through 2003. The graph is highly variable. The first segment shows a direction. The direction is solid. The direction has a first value of 1.97 inches. The second segment shows a direction. The direction is rising. The direction has a last value of 8.9 inches. The graph has the top value of 11 inches. The graph has the minimum value of 0.03 inches.

Select the grade level that best fits the summary above, in your opinion (remember that you can only use each grade level once):
(Dropbox with options: 4th grade, 7th grade, 10th grade, College level)

4. Third summary
A highly variable line diagram which presents the number of annual difference from Seattle's 1899 sea levels in inches and consists of a changing trend composed of a stable trend from 1900 to 1928 followed by a rising trend through 2003 is shown by the picture. A direction which is solid and has an original value of 1.97 inches are given by the first segment. A rising direction which has a last value of 8.9 inches is given by the second segment. The diagram which has the least value of 0.03 inches has the top value of 11 inches.

Select the grade level that best fits the summary above, in your opinion (remember that you can only use each grade level once):
(Dropbox with options: 4th grade, 7th grade, 10th grade, College level)

5. Forth summary
There is an image. The image reveals highly variable a line diagram which presents the number of annual difference from Seattle's 1899 sea levels in inches and consists of a changing trend composed of a stable trend from 1900 to 1928 followed by a rising trend through 2003. The first segment reveals a solid trend. A first value of 1.97 inches is reached by the trend.
The second segment reveals a rising trend which has a last value of 8.9 inches. The diagram which has the minimum value of 0.03 inches has the highest value of 11 inches.

Select the grade level that best fits the summary above, in your opinion (remember that you can only use each grade level once):
(Dropbox with options: 4th grade, 7th grade, 10th grade, College level)

6. Briefly explain what made you choose the summary you associated with 4th grade level
(Text box provided for answer)

7. Briefly explain what made you choose the summary you associated with college level
(Text box provided for answer)

Appendix D

ANALYSIS OF THE RESULTS OF THE MECHANICAL TURK EXPERIMENT

Graph L3:

The desired outcome for this line graph's summaries ordering was: **4 1 3 2**.

The six pairwise relationships are: $\mathbf{4 \succ 1;\ 4 \succ 3;\ 4 \succ 2;\ 1 \succ 3;\ 1 \succ 2;\ 3 \succ 2}$.

Table D-1: Results of applying the pairwise relationship approach to line graph L3.

	Pairwise relationships						
Responses	$\mathbf{4 \succ 1}$	$\mathbf{4 \succ 3}$	$\mathbf{4 \succ 2}$	$\mathbf{1 \succ 3}$	$\mathbf{1 \succ 2}$	$\mathbf{3 \succ 2}$	**count**
(1) 4 1 3 2	x	x	x	x	x	x	6
(2) 4 1 3 2	x	x	x	x	x	x	6
(3) 4 3 2 1	x	x	x			x	4
(4) 3 1 4 2			x		x	x	3
(5) 4 1 3 2	x	x	x	x	x	x	6
(6) 3 1 4 2			x		x	x	3
(8) 2 1 3 4				x			1
Count	4	5	7	4	5	6	**29**

	Pairwise relationships						
Responses	**4 ≻ 1**	**4 ≻ 3**	**4 ≻ 2**	**1 ≻ 3**	**1 ≻ 2**	**3 ≻ 2**	**count**
Prob (%)	57.1%	71.4%	100%	57.1%	71.4%	85.7%	**69%**

The nDCG formulae used for calculation is the following:

$$\mathrm{DCG_p} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)}$$

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p}$$

i	rel i	log2 i	rel i/log2 i	2^rel i - 1/log2 (i+1)		
1	2	0	N/A	1.892789261		
2	4	1	4	6.460148371		
3	3	1.585	1.893	4		
4	1	2	0.5	0		
L3						
Ideal (iDCG)	4	1	3	2	**8.393**	
response 1	4	1	3	2	8.393	1
response 2	4	1	3	2	8.393	1
response 3	4	3	2	1	8.131	0.969
response 4	3	1	4	2	8.131	0.969
response 5	4	1	3	2	8.393	1
response 6	3	1	4	2	8.131	0.969
response 8	2	1	3	4	6.893	0.821
nDCG						**0.961**

Graph L6:

The desired outcome for this line graph's summaries ordering was: **3 4 1 2**. The six pairwise relationships are: **3 ≻ 4; 3 ≻ 1; 3 ≻ 2; 4 ≻ 1; 4 ≻ 2; 1 ≻ 2**.

Table D-2: Results of applying the pairwise relationship approach to line graph L6.

	Pairwise relationships						
Responses	**3 ≻ 4**	**3 ≻ 1**	**3 ≻ 2**	**4 ≻ 1**	**4 ≻ 2**	**1 ≻ 2**	**count**
(1) 4 3 1 2		x	x	x	x	x	5
(2) 3 4 1 2	x	x	x	x	x	x	6
(5) 4 3 2 1		x	x	x	x		4
(7) 4 1 3 2			x	x	x	x	4
(8) 4 3 1 2		x	x	x	x	x	5
Count	1	4	5	5	5	4	**24**
Prob (%)	**20%**	**80%**	**100%**	**100%**	**100%**	**80%**	**80%**

Graph L18:

The desired outcome for this line graph's summaries ordering was: **2 1 4 3**. The six pairwise relationships are: **2 ≻ 1; 2 ≻ 4; 2 ≻ 3; 1 ≻ 4; 1 ≻ 3; 4 ≻ 3**.

Table D-3: Results of applying the pairwise relationship approach to line graph L18.

	Pairwise relationships						
Responses	**2 ≻ 1**	**2 ≻ 4**	**2 ≻ 3**	**1 ≻ 4**	**1 ≻ 3**	**4 ≻ 3**	**count**
(1) 4 3 1 2						x	1
(2) 2 4 1 3	x	x	x		x	x	5
(3) 2 1 4 3	x	x	x	x	x	x	6
(4) 2 1 4 3	x	x	x	x	x	x	6
(5) 3 2 4 1	x	x					2
(8) 4 3 2 1	x					x	2
(9) 3 4 2 1	x						1
Count	6	4	3	2	3	5	**23**
Prob (%)	**85.7%**	**57.1%**	**42.8%**	**28.6%**	**42.8%**	**71.4%**	**54.7%**

Graph L21:

The desired outcome for this line graph's summaries ordering was: **4 1 3 2**. The six pairwise relationships are: **4 ≻ 1; 4 ≻ 3; 4 ≻ 2; 1 ≻ 3; 1 ≻ 2; 3 ≻ 2**.

Table D-4: Results of applying the pairwise relationship approach to line graph L21.

	Pairwise relationships						
Responses	**4 ≻ 1**	**4 ≻ 3**	**4 ≻ 2**	**1 ≻ 3**	**1 ≻ 2**	**3 ≻ 2**	**count**
(1) 2 1 3 4				x			1
(2) 2 1 3 4				x			1
(3) 1 2 4 3		x		x	x		3
(4) 4 2 3 1	x	x	x				3
(5) 4 1 3 2	x	x	x	x	x	x	6
Count	2	3	2	3	2	1	**14**
Prob (%)	**40%**	**60%**	**40%**	**60%**	**40%**	**20%**	**46.6%**

Graph L23:

The desired outcome for this line graph's summaries ordering was: **3 4 1 2**. The six pairwise relationships are: **3 ≻ 4; 3 ≻ 1; 3 ≻ 2; 4 ≻ 1; 4 ≻ 2; 1 ≻ 2**.

Table D-5: Results of applying the pairwise relationship approach to line graph L23.

	Pairwise relationships						
Responses	**3 ≻ 4**	**3 ≻ 1**	**3 ≻ 2**	**4 ≻ 1**	**4 ≻ 2**	**1 ≻ 2**	**count**
(1) 3 4 1 2	x	x	x	x	x	x	6
(2) 4 3 2 1		x	x	x	x		4
(3) 4 3 2 1		x	x	x	x		4
(4) 3 4 1 2	x	x	x	x	x	x	6
(5) 3 4 2 1	x	x	x	x	x		5
(6) 4 2 1 3				x	x		2
(7) 3 2 4 1	x	x	x	x			4
(8) 1 2 3 4	x					x	2
Count	5	6	6	7	6	3	**33**
Prob (%)	**62.5%**	**75%**	**75%**	**87.5%**	**75%**	**37.5%**	**68.7%**

Graph L26:

The desired outcome for this line graph's summaries ordering was: **1 2 3 4**. The six pairwise relationships are: **1 ≻ 2; 1 ≻ 3; 1 ≻ 4; 2 ≻ 3; 2 ≻ 4; 3 ≻ 4**.

Table D-6: Results of applying the pairwise relationship approach to line graph L26.

	Pairwise relationships						
Response	**1 ≻ 2**	**1 ≻ 3**	**1 ≻ 4**	**2 ≻ 3**	**2 ≻ 4**	**3 ≻ 4**	**count**
(1) 1 2 3 4	x	x	x	x	x	x	6
(3) 1 2 3 4	x	x	x	x	x	x	6
(6) 1 2 3 4	x	x	x	x	x	x	6
(7) 1 3 4 2	x	x	x			x	4
Count	4	4	4	3	3	4	**39**
Prob (%)	**100%**	**100%**	**100%**	**75%**	**75%**	**100%**	**91.7%**

Graph L28:

The desired outcome for this line graph's summaries ordering was: **2 1 4 3**. The six pairwise relationships are: **2 ≻ 1; 2 ≻ 4; 2 ≻ 3; 1 ≻ 4; 1 ≻ 3; 4 ≻ 3**.

Table D-7: Results of applying the pairwise relationship approach to line graph L28.

	Pairwise relationships						
Responses	**2 ≻ 1**	**2 ≻ 4**	**2 ≻ 3**	**1 ≻ 4**	**1 ≻ 3**	**4 ≻ 3**	**count**
(2) 1 2 3 4		x	x	x	x		4
(3) 2 1 3 4	x	x	x	x	x		5
(5) 2 3 1 4	x	x	x	x			4
(7) 2 1 3 4	x	x	x	x	x		5
(8) 1 2 4 3		x	x	x	x	x	5
(9) 2 1 4 3	x	x	x	x	x	x	6
Count	4	6	6	6	5	2	**29**
Prob (%)	**66.7%**	**100%**	**100%**	**100%**	**83.3%**	**33.3%**	**80.5%**

Graph L42:

The desired outcome for this line graph's summaries ordering was: **4 1 3 2**. The six pairwise relationships are: **4 ≻ 1; 4 ≻ 3; 4 ≻ 2; 1 ≻ 3; 1 ≻ 2; 3 ≻ 2**.

Table D-8: Results of applying the pairwise relationship approach to line graph L42.

	Pairwise relationships						
Responses	**4 ≻ 1**	**4 ≻ 3**	**4 ≻ 2**	**1 ≻ 3**	**1 ≻ 2**	**3 ≻ 2**	**count**
(4) 1 2 3 4				x	x		2
(6) 4 1 3 2	x	x	x	x	x	x	6
(8) 4 1 2 3	x	x	x	x	x		5
(9) 3 1 4 2			x		x	x	3
Count	2	2	3	3	4	2	**16**
Prob (%)	**50%**	**50%**	**75%**	**75%**	**100%**	**50%**	**66.7%**

Graph L89:

The desired outcome for this line graph's summaries ordering was: **3 4 1 2**. The six pairwise relationships are: **3 ≻ 4; 3 ≻ 1; 3 ≻ 2; 4 ≻ 1; 4 ≻ 2; 1 ≻ 2**.

Table D-9: Results of applying the pairwise relationship approach to line graph L89.

	Pairwise relationships						
Responses	**3 ≻ 4**	**3 ≻ 1**	**3 ≻ 2**	**4 ≻ 1**	**4 ≻ 2**	**1 ≻ 2**	**count**
(2) 4 3 2 1		x	x	x	x		4
(3) 3 4 1 2	x	x	x	x	x	x	6
(4) 4 3 1 2		x	x	x	x	x	5
(6) 4 3 1 2		x	x	x	x	x	5
(8) 4 3 2 1		x	x	x	x		4
Count	1	5	5	5	5	3	**24**
Prob (%)	**20%**	**100%**	**100%**	**100%**	**100%**	**60%**	**80%**

Graph L95:

The desired outcome for this line graph's summaries ordering was: **1 2 3 4**. The six pairwise relationships are: $\mathbf{1 \succ 2;\ 1 \succ 3;\ 1 \succ 4;\ 2 \succ 3;\ 2 \succ 4;\ 3 \succ 4}$.

Table D-10: Results of applying the pairwise relationship approach to line graph L95.

	Pairwise relationship						
Responses	$\mathbf{1 \succ 2}$	$\mathbf{1 \succ 3}$	$\mathbf{1 \succ 4}$	$\mathbf{2 \succ 3}$	$\mathbf{2 \succ 4}$	$\mathbf{3 \succ 4}$	**count**
(2) 1 2 3 4	x	x	x	x	x	x	6
(3) 1 2 3 4	x	x	x	x	x	x	6
(4) 1 2 4 3	x	x	x	x	x		5
(5) 1 2 3 4	x	x	x	x	x	x	6
(6) 1 3 4 2	x	x	x			x	4
(7) 1 2 3 4	x	x	x	x	x	x	6
(9) 1 3 2 4	x	x	x		x	x	5
Count	7	7	7	5	6	6	**38**
Prob (%)	**100%**	**100%**	**100%**	**71.4%**	**85.7%**	**85.7%**	**90.5%**

Appendix E

IRB APPROVALS

RESEARCH OFFICE

210 Hullihen Hall
University of Delaware
Newark, Delaware 19716-1551
Ph: 302/831-2136
Fax: 302/831-2828

DATE: January 25, 2013

TO: Priscilla Moraes, Msc
FROM: University of Delaware IRB

STUDY TITLE: [419550-1] Evaluating Natural Language Generated Text for Describing Information Graphics for Visually Impaired Users

SUBMISSION TYPE: New Project

ACTION: APPROVED
APPROVAL DATE: January 25, 2013
EXPIRATION DATE: January 24, 2014
REVIEW TYPE: Expedited Review

REVIEW CATEGORY: Expedited review category # 7

Thank you for your submission of New Project materials for this research study. The University of Delaware IRB has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a study design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

This submission has received Expedited Review based on the applicable federal regulation.

Please remember that informed consent is a process beginning with a description of the study and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the study via a dialogue between the researcher and research participant. Federal regulations require each participant receive a copy of the signed consent document.

Please note that any revision to previously approved materials must be approved by this office prior to initiation. Please use the appropriate revision forms for this procedure.

All SERIOUS and UNEXPECTED adverse events must be reported to this office. Please use the appropriate adverse event forms for this procedure. All sponsor reporting requirements should also be followed.

Please report all NON-COMPLIANCE issues or COMPLAINTS regarding this study to this office.

Please note that all research records must be retained for a minimum of three years.

Based on the risks, this project requires Continuing Review by this office on an annual basis. Please use the appropriate renewal forms for this procedure.

If you have any questions, please contact Jody-Lynn Berg at (302) 831-1119 or jlberg@udel.edu. Please include your study title and reference number in all correspondence with this office.

RESEARCH OFFICE

210 Hullihen Hall
University of Delaware
Newark, Delaware 19716-1551
Ph: 302/831-2136
Fax: 302/831-2828

DATE: April 10, 2014

TO: Priscilla Moraes
FROM: University of Delaware IRB

STUDY TITLE: [592366-1] Evaluating Natural Language Generated Text Describing Information Graphics for Visually Impaired Users

SUBMISSION TYPE: New Project

ACTION: APPROVED
APPROVAL DATE: April 10, 2014
EXPIRATION DATE: April 9, 2015
REVIEW TYPE: Expedited Review

REVIEW CATEGORY: Expedited review category # *7*

Thank you for your submission of New Project materials for this research study. The University of Delaware IRB has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a study design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

This submission has received Expedited Review based on the applicable federal regulation.

Please remember that informed consent is a process beginning with a description of the study and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the study via a dialogue between the researcher and research participant. Federal regulations require each participant receive a copy of the signed consent document.

Please note that any revision to previously approved materials must be approved by this office prior to initiation. Please use the appropriate revision forms for this procedure.

All SERIOUS and UNEXPECTED adverse events must be reported to this office. Please use the appropriate adverse event forms for this procedure. All sponsor reporting requirements should also be followed.

Please report all NON-COMPLIANCE issues or COMPLAINTS regarding this study to this office.

Please note that all research records must be retained for a minimum of three years.

Based on the risks, this project requires Continuing Review by this office on an annual basis. Please use the appropriate renewal forms for this procedure.

If you have any questions, please contact Nicole Farnese-McFarlane at (302) 831-1119 or nicolefm@udel.edu. Please include your study title and reference number in all correspondence with this office.

RESEARCH OFFICE

210 Hullihen Hall
University of Delaware
Newark, Delaware 19716-1551
Ph: 302/831-2136
Fax: 302/831-2828

DATE: February 26, 2014

TO: Priscilla Moraes
FROM: University of Delaware IRB

STUDY TITLE: [572190-1] Evaluating Natural Language Generated Text Adapted for Different Grade Levels Describing Information Graphics for Visually Impaired Users

SUBMISSION TYPE: New Project

ACTION: DETERMINATION OF EXEMPT STATUS
DECISION DATE: February 26, 2014

REVIEW CATEGORY: Exemption category # *1*

Thank you for your submission of New Project materials for this research study. The University of Delaware IRB has determined this project is EXEMPT FROM IRB REVIEW according to federal regulations.

We will put a copy of this correspondence on file in our office. Please remember to notify us if you make any substantial changes to the project.

If you have any questions, please contact Nicole Famese-McFarlane at (302) 831-1119 or nicolefm@udel.edu. Please include your study title and reference number in all correspondence with this office.