JORGE SASSAKI RESENDE SILVA

AUTOMATICALLY GENERATED HEADERS AS TEXT-SKIMMING MECHANISMS FOR BLIND USERS USING SCREEN READING SOFTWARE IN UNMARKED WEB-BASED TEXTS

LAVRAS – MG

2021

JORGE SASSAKI RESENDE SILVA

AUTOMATICALLY GENERATED HEADERS AS TEXT-SKIMMING MECHANISMS FOR BLIND USERS USING SCREEN READING SOFTWARE IN UNMARKED WEB-BASED TEXTS

Dissertação de mestrado apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de Banco de Dados/Engenharia de Software, para a obtenção do título de Mestre.

Prof. DSc. André Pimenta Freire
Orientador

Prof. DSc. Paula Christina Figueira Cardoso
Coorientadora

LAVRAS – MG
2021

Ficha catalográfica elaborada pela Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Silva, Jorge Sassaki Resende
Automatically generated headers as text-skimming mechanisms for blind users using screen reading software in unmarked web-based texts / Jorge Sassaki Resende Silva. – Lavras : UFLA, 2021.
90 p. : il.

Dissertação (mestrado acadêmico)–Universidade Federal de Lavras, 2021.
Orientador: Prof. DSc. André Pimenta Freire.
Bibliografia.

1. Acessibilidade. 2. Deificiência Visual. 3. Processamento de Linguagem Natural. I. Freire, André Pimenta II. Cardoso, Paula Christina Figueira

JORGE SASSAKI RESENDE SILVA

AUTOMATICALLY GENERATED HEADERS AS TEXT-SKIMMING MECHANISMS FOR BLIND USERS USING SCREEN READING SOFTWARE IN UNMARKED WEB-BASED TEXTS

Dissertação de mestrado apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de Banco de Dados/Engenharia de Software, para a obtenção do título de Mestre.

APROVADA em 23 de novembro de 2021.

Prof. DSc. Thiago Alexandre Salgueiro Pardo USP
Prof. DSc. Natasha Malveira Costa Valentim UFPR

Prof. DSc. André Pimenta Freire
Orientador

Prof. DSc. Paula Christina Figueira Cardoso
Coorientadora

LAVRAS – MG
2021

A todos que, em tempos obscurecidos por negacionismo, insistem na Ciência.

AGRADECIMENTOS

Agradeço à Universidade Federal de Lavras – UFLA e ao Departamento de Ciência da Computação – DCC pela estrutura oferecida para realização da pesquisa. Também pelo apoio financeiro fornecido através de bolsas de monitoria.

Agradeço aos professor André Pimenta Freire e à professora Paula Christina Figueira Cardoso por todo o apoio, orientação, paciência e disposição para ajudar em todas as etapas deste trabalho.

Agradeço a todos os voluntários que participaram do estudo de usuário, pela disponibilidade e boa vontade em contribuir com esta pesquisa.

Agradeço a todos que me apoiaram e ajudaram no decorrer do programa de Mestrado, em especial meus pais Silvio e Helena.

Agradeço ao meu amigo Augusto, que foi meu colega na Graduação e Mestrado, pela companhia nas caminhadas na UFLA e nas horas em laboratório.

Agradeço a minha psicóloga Aline que me ajudou a não jogar tudo pro alto e sair correndo.

Agradeço a todos que me fizeram companhia on-line durante os tempos difíceis de isolamento social da pandemia do coronavírus, em especial à Buh que passou muitas horas comigo durante a escrita deste trabalho via Discord.

Despite everything, it's still you.
(Toby Fox, Undertale, 2015)

RESUMO

A evolução da Internet ampliou a disponibilidade de informação e de conteúdos didáticos na Web ampliando o acesso ao conhecimento. Mesmo assim, as pessoas com deficiência enfrentam problemas de acessibilidade. Neste trabalho, focamos em pessoas com deficiência visual e seus problemas na leitura de documentos longos. Mesmo com leitores de tela, buscar informações na Web pode ser uma tarefa exaustiva. Faltam opções para leitura rápida de grandes quantidades de texto, uma vez que leitores de tela lêem o texto de forma sequencial. Os comandos disponíveis por meio de atalhos ajudam os usuários a criar estratégias na busca de informações, como aumento da velocidade de leitura e uso de marcação de página da web para navegar. Mas muitas vezes eles têm um impacto na carga cognitiva por causa da atenção necessária para entender e lembrar informações em alta velocidade. Navegar dentro de textos também é um problema, estudos mostram que um problema comum é ter que mapear mentalmente as informações do texto. Pensando nisso, realizamos um mapeamento sistemático para reunir trabalhos anteriores que visavam ajudar os usuários de leitores de tela no uso da Web, e os categorizamos com base em suas estratégias. Os métodos encontrados foram Filtragem de Conteúdo, Redução de Texto, Navegação, Fala Simultânea, Visão Geral Auditiva e Sistemas de Recomendação. Observamos uma falta de trabalhos com topicalização e seus efeitos na navegação de cegos. Assim, o objetivo deste trabalho foi propor um algoritmo para gerar cabeçalhos automaticamente para auxiliar usuários na busca de informação. O algoritmo foi dividido em duas tarefas: segmentar um documento por tópicos e rotular estes segmentos. Adaptamos o algoritmo de segmentação C99 para usar o BERT e observamos melhora nas taxas de erro para textos longos. Em seguida, foi implementado um algoritmo de rotulagem baseado em palavras-chave, os rótulos são feitos das palavras mais repetidas no documento. Para testar o algoritmo, conduzimos um estudo de usuário com 8 participants e um protótipo composto de 4 textos pré-processados de 720-1131 palavras. Os usuários tinham que responder a conjuntos de perguntas com base nas informações desses textos, para comparação em dois cenários: um com cabeçalhos gerados automaticamente e outro sem. Medimos o tempo gasto em cada texto e a carga cognitiva que os participantes sentiram ao realizar as tarefas. Uma entrevista pós-teste também foi conduzida para coletar feedback. Nossa análise não pôde confirmar nossa hipótese com alta significância devido à pequena amostra de voluntários, mas as entrevistas indicaram que os usuários se beneficiaram com a ferramenta proposta. Seja ajudando a navegar dentro do texto ou reencontrando informações, os participantes concordaram que gostariam de ter essa ferramenta em seus leitores de tela. Com este trabalho, fornecemos implicações de design e alternativas para implementar um plugin para leitor de tela.

Palavras-chave: Acessibilidade. Deficiência visual. Processamento de linguagem natural

ABSTRACT

The evolution in information access caused by the Internet has expanded access to information for everyone. Consequently, the availability of information sources and teaching content on the Web has broadened access to knowledge. Still, people with disabilities face accessibility barriers. In this work, we focus on people with visual impairment and their problems when reading long documents. Even with assistive technologies such as screen readers, consulting the Web when seeking information can be exhausting for this group. They lack options for speed reading or skimming large amounts of text since these tools read texts aloud in sequence. The commands available in screen readers through shortcuts helped users create strategies when searching for information, such as increased reading speed and using webpage markup to navigate (*e.g.* headers or paragraphs). However, they often come at the cost of the cognitive load caused by the attention needed to understand and remember the information at high speeds. Navigating inside texts also proved to be a problem. Studies have shown that a common complaint of visually impaired people is having to create a text map to re-find information mentally. To counter these problems, we have conducted a systematic mapping to gather previous work that aimed to help screen reader users when using the Web and categorize them based on their proposed approach. The methods encountered were Content Filtering, Text Reduction, Navigation, Concurrent Speech, Auditory Overview and Recommendation Systems. Based on this analysis, we observed a lack of topicalization techniques and their effects on navigation for blind people. Thus, the goal of this work was to propose an algorithm to generate headers aiming to help users in information-seeking tasks automatically. The algorithm was divided into two tasks: segmenting a document into topic segments and labelling a text segment. We adapted the C99 segmenting algorithm to use BERT and improved error rates for long texts. Then, the study followed with the implementation of a labelling algorithm based on keywords, and labels are made of words from the text segment that were ranked according to repetition. We conducted a user study with 8 participants and a prototype composed of preprocessed texts 720-1131 words long to test the algorithm. Users had to answer questions based on the information in these texts for comparison in two scenarios: one with automatically generated headers and the other without. We measured the time taken in each text and the cognitive load participants perceived while completing it. A post-test interview was also conducted to gather feedback. Our analysis could not confirm our hypothesis significantly due to a small volunteer sample, but interviews indicated users benefited from the proposed tool. Either by helping navigate inside the text, or re-finding information, participants agreed they would like to have this tool in their screen readers. With this work, we provide design implications and alternatives to implement a plugin for the screen reader.

Keywords: Accessibility. Blind people. Natural Language Processing

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

The Internet popularization has made it increasingly relevant to society. It has become one of the primary sources of information and is commonly used in contexts such as communication, social networking, news reading and information research. Due to the massive flow of information on the network, users can quickly feel overwhelmed by news websites or social networks. This fact is aggravated when using screen reader software for computer access for blind people.

Screen readers need to narrate information contained in computers linearly, reading each line of text word for word employing voice output. This reading process can be very inefficient and lead to unnecessary cognitive load for visually impaired people depending on the structure of the content read (GIRAUD; THÉROUANNE; STEINER, 2018). This problem can become worse in information seeking tasks in which users often need to search for specific information and find it in the middle of long documents. Without the option of using visual memory to quickly find information, they often need to remember any information that might seem useful. If the user forgets something, they need to read most of the text again to rediscover the information (GHAHARI et al., 2012).

Sighted users can reduce the information overload caused by the reading of excessive amounts of online content using skimming methods. These methods range from identifying elements of interest on a page by glancing at its structure and visual cues and using visual memory to remember the path taken to find specific information. Websites are commonly designed with sighted users in mind, with many visual cues to help the reader quickly identify sections of interest and present content comfortably. However, the same websites are often not accessible to visually-impaired users, either because their structure is not suitable for keyboard navigation or the lack of alternative text for images or topics to organise and segment long texts, among other factors.

Another issue screen reader users face frequently is time. As the reading of the text needs to be done entirely without jumping words or sentences, it becomes an aggravating factor for the user's tiredness. There are limited options available for saving time while using screen readers. Users may use keyboard shortcuts to skip content (commonly used to skip irrelevant sections such as menus or page headers), navigate using alternative methods (*e.g.* header or link listings), or accelerate the text reading speed. Visually impaired people often do not skip content inside long texts because they feel insecure about losing important information. Increasing the rate at which content is read is a widely used option. However, it still is not the ideal as increasing the flow of information can increase the cognitive load required and demands a higher level of attention from the user (GADDE; BOLCHINI, 2014).

To keep up with the fast-growing amount of information shared online and the significant frequency with which it is updated, research has been done to develop new technologies and techniques to enable visually impaired people to access the Internet more comfortably. It is crucial to ensure that modern technologies (*e.g.* interactive controls in virtual classrooms), that rapidly shape virtual space, do not create even more significant barriers (such as controls that work exclusively with a mouse) to inclusive information access.

An efficient way to minimize the cognitive load of screen reader users is to reduce the information to be carried by the software. According to Rayner et al. (2016), skimming is a reasonable strategy to cope with the overwhelming amount of text we have to read, given we are willing to accept a trade-off between speed and accuracy. Ahmed et al. (2012) proposed a method for visually impaired people to skim a text. Skimming is a strategy commonly used by sighted people to quickly understand the subject of a given text while searching for specific information. One way of skimming is to skip words in the text, reading only those that catch the reader's attention, reducing the time spent in the text without becoming lost. In the non-visual approach of Ahmed *et al.*, an automatic summarization method was proposed to reduce the content of a text by summarizing every sentence separately. This summary was designed to be independent of a corpus or word frequency to suit better the nature of the internet, which often broadcasts information through short texts.

An efficient navigation method to be used with the generated summary is necessary, aiming at greater comfort of screen reader users in their interactions with computers. Besides, there is a need for a summarization strategy that assists users in reading documents found on the web, such as longer articles and newspaper columns, digitally published menus, notices, regulations, and other long texts. Reducing text alone is not enough to ensure ease of access to information. It is also necessary to provide means of navigating from the summary text or automatically-generated headers to the original text quickly so that the user has access to more details if needed.

Another possible way to help screen reader users when navigating through long text is by generating headers for each topic of the document. This way, even if the document has large text blocks, the user can know what to expect from each part of the text and use the headers as anchors to navigate through the text. Although the idea that organizing information in smaller segments increases readability is not new (TOMBAUGH; LICKORISH; WRIGHT, 1987), we did not find applications of text segmentation and topic labelling for screen readers.

Despite advancements in the exploration of automatic approaches to enable text skimming for visually-impaired users (AHMED et al., 2012; GADDE; BOLCHINI, 2014; GIRAUD; THÉROUANNE; STEINER, 2018; ALVES; CARDOSO; FREIRE, 2018), there is still limited availability of such solutions to visually impaired users. Further to the lack of availability of such tools in free screen reader technologies, there are no such tools available for Brazilian Portuguese.

This research involved a systematic mapping (Appendix A) to provide an overview of previous studies in skim-reading techniques for screen readers. The works found in this review were categorized according to the approach used to enable visually impaired users to skim documents. This categorization helped in finding a gap to propose a new tool for aiding screen reader users.

Thinking about easing the difficulties faced by screen reader users when browsing texts on the Web, in this work we take a step forward by exploring the automatic generation of headers coupled to the screen reader to provide a skimming strategy alternative for visually impaired people in the Brazilian context. The present work describes an algorithm for subdividing expository text into subtopics, relying on the idea that multiple sub-topical discussion occur during the discussion of one or multiple main topics (HEARST, 1997). This work focused on the Portuguese Language.

We conducted an evaluation of a proposed heading generating algorithm by conducting a user study with visually-impaired volunteers. The participants were regular screen reader users, with Portuguese as their native language. We expected that with the aid of automatically generated headers, users would save time and feel less tired during information seeking tasks.

The goal of this research was to investigate how to provide a better usability for screen reader users in reading texts on the Web. To achieve this we intended to answer the following question:

How can automatic header generation with topic segmentation algorithms be used to improve screen reader usability in information seeking tasks in texts in Portuguese?

This project's **main objective** was to propose an algorithm to automatically generate headers aiming to help visually-impaired users in information-seeking tasks. We expected to observe an improvement in information-seeking task performance.

The specific objectives of this project were:

a) To develop topic segmentation and topic labelling algorithms to automatically generate headers;
b) To evaluate the implemented algorithm with screen-reader users.

This document is organized as follows. Chapter 1 contextualizes the problem and describes the objectives of this project. Chapter 2 defines the main concepts used in this work, including basic concepts about accessibility, text skimming, neural embeddings, topic segmentation, topic labelling and related work. Chapter 3 describes the methodology used for the implementation of the proposed automatic header generation algorithm. Chapter 4 describes the user study methodology to evaluate our proposed tool and presents its results. In Chapter 5 we discuss the algorithm performance and implications encountered in the user study, as well as the viability of actually implementing this algorithm as a screen reader plugin and acknowledge the limitations of this project. Chapter 6 gathers the final considerations of this work.

2 THEORETICAL BACKGROUND

This study applies algorithms to automatically generate headers based on topics, resulting in a header list to help screen reader users in information-seeking tasks. Thus, it involves aspects of Accessibility and Natural Language Processing. This Chapter defines key concepts used through this study in those two areas and presents related work concerning ways of providing automatic skim-reading techniques for visually-impaired users.

2.1 Web accessibility for visually impaired people

According to the World Health Organization (World Health Organization; World Bank, 2011), about 15% of the world's population live with disability. In Brazil, the percentage of the population with disability is about 24%, according with the latest census available (IBGE, 2011). In December 13th 2006 the United Nations created the Convention on the Rights of Persons with Disabilities with the purpose of promoting, protecting and ensuring the full and equal enjoyment of all human rights and fundamental freedoms by all persons with disabilities, and promoting respect for their inherent dignity. This convention established the term "people with disabilities" including those who have long-term physical, mental, intellectual or sensory impairments, which may pose barriers that hinder their full and effective participation in society on an equal basis with others. Two years later, this convention was incorporated in Brazil's legislation by the federal decree number 186 from 2008 (BRASIL, 2008).

According to the Brazilian Law for Inclusion of Persons with Disabilities (LBI), a person with disability is someone with a long-term impairment of a physical, mental, intellectual or sensory cause, which, in interaction with one or more barriers, can obstruct their full and effective participation in society on equal terms as other people.

Accessibility is essential to overcome barriers faced by people with disabilities, enabling them to live independently and to fully participate in society and all aspects of life with autonomy. It is also worth noting that, although disabilities and accessibility are often closely related, the latter is also applicable for other people who do not have disabilities *per se*, such as older people, people with temporary limitations (*e.g.* post-op patients) or with reduced mobility (*e.g.* pregnant women).

This study focuses on visually-impaired people who often use assistive technologies in their daily routines. According to the LBI (BRASIL, 2015), assistive technologies can be defined as "any item, equipment part, or product, acquired in commerce or adapted or modified, used to augment, maintain or improve the functional capacity of people with disabilities". Assistive technologies are powerful tools to increase independence and to better participation of its users in society (RIBEIRO; FILHO; LEITE, 2016).

A screen reader is a particularly useful assistive technology for people with visual impairment. It provides information about elements displayed on the computer screen through speech synthesis. The user can use keyboard commands to navigate interfaces' different elements while using the screen reader. It interacts with the operating system, capturing the information presented in the form of text and transforming it in audio through a speech synthesizer. Not only do they read the text present on web pages, but also information tagged with HTML tags such as headers, links, paragraphs, and buttons. These features are useful to convey additional information that would otherwise be only visible (*e.g.* sighted users know when a piece of text is a link by its underline and blue convention).

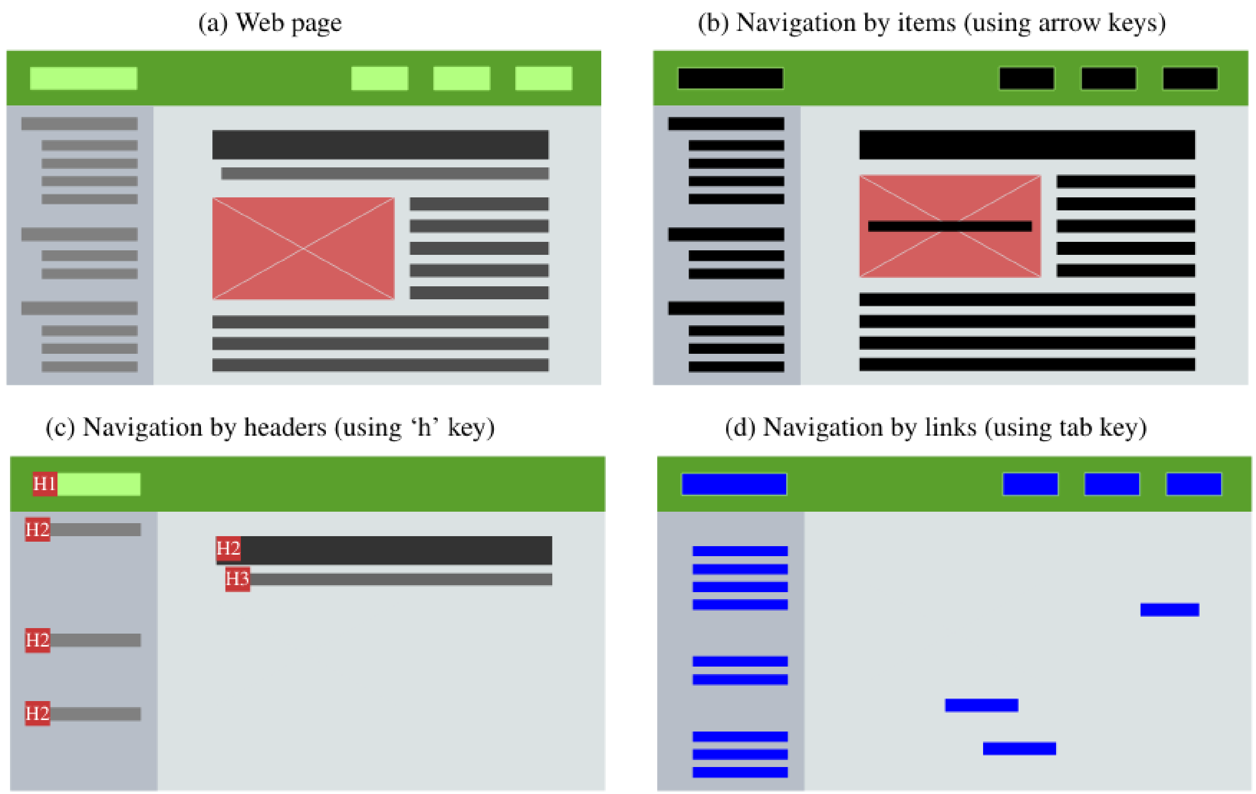
Examples of screen readers include: NVDA (NonVisual Desktop Access)[1], available for free in Windows; JAWS (Job Access With Speech)[2], a commercial software available for Windows; ChromeVox[3], free software available as an extension of Google Chrome's desktop version; Orca[4], open source and free software available in Linux; TalkBack[5], default in Android; VoiceOver[6], built in iOS and MacOS.

Users have different ways of navigating a web page while using a screen reader. Figure 2.1 illustrates the options available in most screen readers. Figure 2.1.a is a given web page. Figure 2.1.b is the most basic navigation method for screen readers. The user navigates through the items tree from the HTML page. This strategy is also the most verbose, considering it reads all possible content in the document. Figure 2.1.c shows the header navigation, where the user listens only to the text tagged as headers in the HTML document (`<h1>` through `<h6>`). Figure 2.1.d shows another common way of navigating when searching for information, which is by using the links on the page.

The linearity of screen reader narration and the large amount of information present in long texts can cause cognitive overload (BUCY; NEWHAGEN, 2004; SUNDAR, 2004), negatively affecting information absorption and causing frustration for the user. **Cognitive load** refers to the memory resources used whilst doing activities (WARNICK et al., 2005). It is divided into three categories: intrinsic, which is associated with the inherent difficulty of a specific topic; the extraneous, which is related to the way information is

[1] Available at <https://www.nvaccess.org>
[2] Available at <https://www.freedomscientific.com/products/software/jaws/>
[3] Available at <https://chrome.google.com/webstore/detail/chromevox-classic-extensi/kgejglhpjiefppelpmljglcjbhoiplfn>
[4] Available at <https://help.gnome.org/users/orca/stable/>
[5] Available at <https://play.google.com/store/apps/details?id=com.google.android.marvin.talkback>
[6] Available at <https://www.apple.com/accessibility/mac/vision/>

Figure 2.1 – Different screen reader navigation options



Source: from the author (2021)

presented; and the germane, which results from the effort to permanently store acquired information in memory (CHANDLER; SWELLER, 1991).

The focus of the present study will be on the extraneous cognitive load, seeking to find a way to promote more comfortable access to information using screen readers by presenting the information to users to reduce the demand for cognitive resources.

2.2 Web Navigation Strategies for Visually Impaired People

When navigating the Internet, people do not usually read all content presented to them. A good example is Banner Blindness (PERNICE, 2018), a well-known problem in web advertising. Users often completely ignore banner ads when browsing on-line and this behaviour is unconscious, people automatically do not pay attention to ad-like content. Users also recognize easily what areas of a web-page are menus, what can be ignored and what is the content they are looking for. Either by patterns, color scheme, difference in font-size, these cues help us identify what we should be paying attention for. Even relevant text is often not read in its entirety because users apply Text Skimming to save time.

Text Skimming is a technique to identify the main topics of a text quickly. People commonly use this strategy when they have large amounts of material to read in a limited amount of time. This process is usually done three to four times faster than regular reading. Its goal is to quickly move one's eyes through the text, looking for specific words, a piece of information, or to get a general idea of the text content. Due to the increased speed and the high number of ignored words, there is a trade-off between speed and comprehension accuracy. For example, the main ideas of a text can still be perceived, but it is harder to make inferences (RAYNER et al., 2016). Skimming is useful when the reader is seeking specific information instead of fully understanding the text.

Effective skimmers adopt an adaptive reading technique, where they scan a text for headings, keywords, and paragraph structure to locate potentially relevant information and, then, reading more carefully when such regions are found. Effective skimming means making sensible decisions about which parts of a text to select for more careful reading when faced with time pressure (RAYNER et al., 2016).

Regarding the effectiveness of text skimming for increasing comprehension while decreasing time, Duggan and Payne (2009) conducted a set of studies that demonstrated an increase in understanding of a text in a limited amount of time. Skimming did not aid memory for less important details, and it did not facilitate inferences about information from the text. Nevertheless, it did allow for better memory regarding the most important ideas from texts. It is also worth noting that this was only observable for long texts. Previous studies (MASSON, 1982; CARVER, 1984; DYSON; HASELGROVE, 2000) with smaller texts did not show relevant advantage in skimming.

While text skimming is widely used by sighted people, this strategy is unavailable for visually impaired users since they cannot look forward or diagonally inside a large text. Since the Internet has become a widely used source of information, and an important tool for daily tasks, visually impaired people had to adapt their use of screen readers for the ever growing digital life. Hence, different strategies are applied to save time when navigating the web for instance: techniques to mentally map an often used website, quickly finding important information, ignoring uninteresting content and avoiding barriers caused by bad accessibility.

According to Borodin et al. (2010), Web browsing for visually impaired people can become a big challenge, and still cause frustration due to potential accessibility and usability problems. Moreno et al. (2018) mention that even with the creation of guidelines like WCAG 2.0, most obstacles faced by people with disabilities are not covered by these guidelines. Ferri and Favalli (2018) state in their study that the advance of technology, although bringing several positive changes in some scenarios is still a cause of exclusion and non-inclusion.

Considering the challenges faced by visually impaired people, Borodin et al. (2010) and Power et al. (2013) studied different strategies used to overcome the barriers found in web accessibility when using computers. Even though it is not the focus of this project, it is worth noting that similar research has been conducted in the context of smartphones usage (JAIN; DIWAKAR; SWAMINATHAN, 2021).

According to Borodin et al. (2010), expert users often speed up the speech rate with which screen readers narrates the content. They prefer older formative speech synthesizers whose quality does not degrade as much at high speech rates as newer speech synthesizers. People who grew up listening to synthesized speech can reach speeds up to 500 words per minute.

Another strategy commonly used is to get a general overview of a page by navigating all headings. If the information of interest cannot be found, they return to the beginning to read through the entire page. For frequently visited websites, experienced users tend to remember the order of headings and other landmarks, which helps to quickly navigate to the desired page or section. Power et al. (2013) corroborates this strategy with two types of strategy they identified. Both *Navigation* and *Discovery* strategies relate to a user navigating inside a website gaining an overview of its overall structure by probing headings and page items to figure out key sections.

Skipping irrelevant content is another way to speed up the search for important content. To this end, users can navigate through headings, keyword search or skipping to non-linked content (BORODIN et al., 2010; POWER et al., 2013). With heading navigation, users can jump through different levels of headings (`<H1>-<H6>`). Keyword search helps users when they know what they are looking for *e.g.* what words occur in the title of an article. Skipping to non-linked content in paragraphs is also useful according to, since irrelevant content often contains many links.

Other works that further corroborates with aforementioned findings are the (EVERIS, 2020; WEBAIM, 2021) surveys. WebAim's survey had 1600 participants from around the world but mostly from Europe (23.5%) and North America (57.7%). This work shows most participants (67.7%) navigate through paragraphs as their main navigation strategy followed by the Find feature (13.9%). Table 2.1 shows all data obtained for strategies when navigating web-pages.

Everis Brasil (EVERIS, 2020) did a survey to understand how Brazilian users interact with screen readers when navigating web-pages. This survey was heavily based on WebAIM (2019). The study collected

Table 2.1 – Strategies used by users worldwide

Strategy	# of respondents	% of respondents
Navigate through the headings on the page	1047	67,7
Use the Find feature	215	13,9
Read through the page	126	8,1
Navigate through the links of the page	110	7,1
Navigate through the landmarks/regions of the page	49	3,2

Source: WebAIM (2021)

answers from 427 participants with most from southeast (52.7%) and northeast (21.1%). Table 2.2 shows the results for different strategies used when navigating web-pages. The difference between this study and WebAim is perceptible. While worldwide heading navigation is widely the most used strategy, in Brazil it is only the second (27.9%), behind *read through page* (39.6%). Despite of *Find feature* being the second most used strategy worldwide, in Brazil it is only the fourth (6.6%).

Table 2.2 – Strategies used by Brazilian users

Strategy	# of respondents	% of respondents
Read through the page	169	39,6
Navigate through the headings on the page	119	27,9
Navigate through the links of the page	104	24,4
Use the Find feature	28	6,6
Navigate through the landmarks/regions of the page	7	1,6

Source: Everis (2020)

This difference might be explained by the fact that many visually impaired people in Brazil were taught how to use screen readers with old software that did not have more advanced features. The fact that the scenario in Brazil is so different from the North America and Europe justifies adapting studies for the Brazilian context.

Another strategy for skipping irrelevant content are skip links. Accessibility guidelines require a skip-to-main-content link often placed hidden at the start of a web-page to allow screen reader users too skip navigation menus and start reading the main content of a page. However, since this link is often hidden, developers often fail to maintain skip links. According to both Power et al. (2013) and Borodin et al. (2010), users can be reluctant to use skip links due to lack of control of where the link would take them and also fear of missing information. According to Everis, only 19,4% of users in Brazil always use skip links, while 26.9% uses sometimes and almost 10% never uses. This trend is the same worldwide. According to WebAIM, only 16.8% of users always use skip links while 28.4% uses sometimes and 14.4% never uses.

Gadde and Bolchini (2014) further corroborates this finding by showing an example of bad usage of skip links. They were trying to find navigation issues while searching for products on Amazon. The skip link in the search page would jump to product listings, making so users who used it would not realize the existence of a faceted search system.

It is worth noting that, according to Borodin et al. (2010), even knowing all aforementioned strategies, the choice for using each of them is completely personal and might vary according to context (e.g. reading headlines on a news website might ensue different strategies than e-commerce).

In this study we focus on headings navigation strategy, considering its popularity worldwide and that, even in Brazil, it is still the most popular alternative to reading the whole page. We apply this strategy in large texts to assist users in information seeking tasks. To achieve this, the text needs to first be segmented in different topics. Then, each topic is labeled to automatically generate a heading with which users can navigate inside the text.

2.3 Characterization of the text organization

In this Section we define textual structure concepts used throughout this work. Well-written text can be organized in topical segments describing different aspects of a main subject (HEARST, 1997). For example, one text about Princess Isabel can describe aspects about her wedding, public life, etc. Topical segments can be paragraphs or sentences. Identify the location where the author changes the attention is a challenge.

On the Web, when it comes to long texts, subtopic structure is often marked by headings or subheadings. This facilitates screen reader users to locate themselves inside a piece of text. However this is not always the case. Sometimes texts are not well-structured and might consist of long sequences of paragraphs with little structure demarcation. This can cause fatigue for blind users (GHAHARI et al., 2012) making it difficult to navigate between different subtopics and finding relevant information. Hence, subtopical segmentation can improve usability by allowing one to navigate between multiple topics inside a long text. Tombaugh, Lickorish and Wright (1987) observed that organizing contents of long texts in smaller fragments, or windows, provides spatial cues about locations of portions of previously read texts, aiding in recall and relocation of information already read.

We focus on expository text which are text that explicitly teaches or explains about a topic, since it is better suited to the target applications of information retrieval. The algorithm relies on a sentence-level discourse structure based on **topic shift**. The goal is to partition text into contiguous non-overlapping subtopic segments. The term subtopic here is meant to signify pieces of text *about* something.

According to Brown et al. (1983) *apud* Hearst (1997) the task of attempting to define 'what a topic is' is difficult, so it is easier to focus on describing what can be recognized as topic shift. That is, between two contiguous pieces of discourse which are intuitively considered to have two different 'topics', there should be a point where the shift from one topic to the other is marked. Characterizing this topic shift means finding basis for dividing pieces of discourse in smaller chunks, each with its separate topics.

In this work we focus on identifying topical shift and label them automatically. In the next section we highlight the main related works in topical segmentation.

2.4 Related works in topic segmentation

Most documents address different topics or different aspects of the same subject. Topic segmentation is the task of finding boundaries between subjects in a text relying on clues for a topic shift inside a document (REYNAR, 1998). Text segmenters can be divided into two main groups based on what information is used to detect topic shift: *endogenous* and *exogenous* groups (NAILI; CHAIBI; Ben Ghezala, 2017). The first is based on lexical information extracted from the text, algorithms in this group usually rely on lexical repetition. The second group which uses information outside the text, such as word embedding models.

The *endogeneous* works, which are based on lexical information, that can be highlighted are TextTiling (HEARST, 1997) and C99 (CHOI, 2000).

Though it is possible to use linguistic clues for segmenting topics, most studies in the literature (CHOI, 2000; HEARST, 1997; KOZIMA, 1993) were motivated by the observation that texts are composed of logically ordered sets of related words and usually contains repeating or similar words. Hints for topic boundaries can include repetition of character sequences, word patterns repetition, word frequency, collocations, synonyms, and linguistic cues to measure the similarity between sentences.

In an early research study, Hearst (1997) introduced *TextTiling*, an algorithm that divided a single document into topic blocks composed of multiple paragraphs based on lexical co-occurrence patterns. The similarities between paragraphs were measured based on the word sets present in them. If the word set changed between neighboring paragraphs, it indicates a topic shift. Salton et al. (1997) used term weight methods, such as TF-IDF, to measure the similarity between paragraphs. Two paragraphs are linked when similarity is above a threshold, forming a relationship map. A series of linked paragraphs indicated topics. In addition, Choi (2000) used word frequency to compare sentences and create a similarity matrix representing a document. A maximisation algorithm (REYNAR, 1998) is applied to divide the document into clusters of sentences with high similarity, the topic segments.

Another important contribution of Choi (2000) was the creation of the Choi dataset for text segmentation. The Choi dataset is used to test whether a segmentation algorithm can correctly guess topic boundaries From a subset of the Brown corpus (the ca**.pos and cj**.pos sets), it chooses 10 texts at random and extracts *n* consecutive sentences of each to concatenate and form a new document. The number of sentences *n* taken from each document is chosen uniformly at random within a range specified by the subset id (3–5, 6–8, 9–11, 3–11) identified by min-max number of sentences, *i.e.* 3-5 can have segments of size ranging from 3 to 5 sentences. 3-5, 6-8 and 9-11 each have 100 example documents, and the 3-11 has 400 documents. The dataset can be obtained from an archived version of the C99 segmentation code release[7].

The *exogenous* group can make use of many models such as latent semantic analysis (CHOI; WIEMER-HASTINGS; MOORE, 2001; MISRA et al., 2011; ALEMI; GINSPARG, 2015), generative Bayesian model (EISENSTEIN; BARZILAY, 2008) and word embeddings (NAILI; CHAIBI; Ben Ghezala, 2017; SOLBIATI et al., 2021).

Word embeddings are techniques to represent words by projecting them in a space using numerical vectors. They are applied in many tasks, one of them being text segmentation. In this case, the segmentation task is based on the similarity between vectors generated by the word embedding models. If the similarity between consecutive sentences decreases, it indicates a possible topic-shift. Since they are the focus of this work, we will emphasize the models and how they can be applied to text segmentation in the next section.

Latent Semantic Analysis (LSA), is a widely used methods for word meaning representation. Briefly stated, LSA rests on the thesis that analyzing the contexts in which words occur permits an estimation of their similarity in meaning (DEERWESTER et al., 1990; LANDAUER; DUMAIS, 1997). LSA takes as input a training corpus and constructs a word by document co-occurrence matrix. Typically, normalization is applied to decrease the weight of common words that add little informational value in the words-documents matrix, usually tf-idf. Finally, a dimensionality reduction is implemented by a truncated Singular Value Decomposition (SVD), which projects every word in a subspace of a predefined number of dimensions. This reduces the size of the document axis from the matrix while preserving the similarity structure in the word axis. Once the vectorial representation of words is obtained, the semantic similarity between two terms is typically computed by the **cosine similarity**, that is, the cosine of the angle between them.

[7] http://web.archive.org/web/20010422042459/http://www.cs.man.ac.uk/ choif/software/C99-1.2-release.tgz

An intrinsic difference between LSA and word embeddings is that while LSA is a counter based model, word embeddings are a prediction-based model. Altszyler et al. (2016) compared LSA with Word2Vec in small corpus and observed that while prediction-based models perform better for large corporas (BARONI; DINU; KRUSZEWSKI, 2014), a counter based model can be better suited in situations data is scarce.

Choi, Wiemer-Hastings and Moore (2001) claimed that it was possible to improve the inter-sentence similarities index from their original segmenting algorithm (C99) by taking into account the semantic proximities between words estimated on the basis of LSA. Proximity between any two sentences (or any other textual units), even if these sentences were not present in the original corpus, can be estimated by computing a vector for each of these units—which corresponds to the weighted sum of the vectors of the words that compose it—and then by computing the cosine between these vectors. Choi, Wiemer-Hastings and Moore (2001) have shown that using this procedure to compute the inter-sentence similarities results in the previous version of the algorithm (based solely on word repetition) being outperformed.

Eisenstein and Barzilay (2008) placed lexical cohesion in a probabilistic context, modeling the words in each topic segment as draws from a multinominal language model. By using Bayesian inference[8], Eisenstein and Barzilay could incorporate additional features such as cue phrases.

In this study, we based ourselves on a more recent algorithm proposed by Naili, Chaibi and Ben Ghezala (2017), which uses word embeddings to segment text. They aimed to improve Choi's topic segmentation algorithm (CHOI, 2000) by adding semantics with Word2Vec in the same fashion as Choi, Wiemer-Hastings and Moore (2001)—substituting the count based similarity calculation with cosine similarity. For each sentence in a document, a similarity matrix is calculated using equation (2.1). A rank mask is then applied to the similarity matrix; based on this, the algorithm builds a rank matrix where each cell is the number of neighboring elements that belong to the rank mask with lower values. Then, Reynar's maximisation algorithm (REYNAR, 1994) is used to detect topic boundaries. We follow this trend now applying BERT to C99.

$$Sim(S_1,S_2) = \frac{\sum_{t_i \in S_1 \cap M} \sum_{t_j \in S_2 \cap M} (Ft_i Ft_j cos(Vt_i, Vt_j))}{\sum_{t_i \in S_1} (Ft_i) \sum t_j \in S_2(Ft_j)} \tag{2.1}$$

With S_1 and S_2 correspond to sentences 1 and 2; M corresponds to the word embedding model; Ft_i and Ft_j correspond to the frequency of terms t_i and t_j; Vt_i and Vt_j correspond to the vectors of t_i and t_j in M.

[8] Bayesian inference is a statistical inference method which updates the probability of an event, based on prior knowledge of conditions that might be related to the event.

2.4.1 Word Embeddings

Word embeddings are a numerical representation of words that maps each word in a corpus vocabulary to a set of real-valued vectors in a vector space model. The value vector for representing each word from a corpus vocabulary is learned through supervised techniques such as sentiment analysis and document classification, or unsupervised techniques such as statistical analysis of documents. These models were a breakthrough in NLP because it is possible to preserve semantic, syntatic and contextual meaning of a word based on its use in sentences while still being scalable. Many works used word embedding for text segmentation (NAILI; HABACHA; GHEZALA, 2018; NAILI; CHAIBI; Ben Ghezala, 2017; ALEMI; GINSPARG, 2015; CHOI; WIEMER-HASTINGS; MOORE, 2001; MISRA et al., 2011; SOLBIATI et al., 2021).

Figure 2.2 – One-hot embedding for the sentence “we are meeting online.”

We are meeting online.

We	{ 1 0 0 0 }
are	{ 0 1 0 0 }
meeting	{ 0 0 1 0 }
online	{ 0 0 0 1 }

Source: from the author (2021)

Previous models such as one-hot encoding were sparse and did not preserve context. One-hot encoding creates a vector for each word in a sentence, the size of each vector equals the number of words. All values in the vector are zeros except for one. As an example, take the sentence *we are meeting online* in Figure 2.2. There are 4 different words, therefore the size of each vector is 4. Even in such a small sentence the sparsity is perceptible. If we use longer sentences or add new ones to our encoding, it is easy to imagine a scalability issue. This is illustrated in Figure 2.3, where we take the previous encoding and add the sentence “We are not meeting in person.”. Because each word is simply a vector of zeros and a 1, we also have no meaning information, it’s only a numerical representation. This is the reason word embeddings are important for NLP, enabling the representation of words in a dense, scalable and meaningful model.

In word embeddings, words that have a similar semantic or contextual meaning also have a similar vector representation while each word in the vocabulary has a unique vector representation. The relation between words is also preserved such as different verb tenses and gender, allowing researchers to apply mathematical properties to words. For example, since semantically or contextually related words have similar representations in a vector space model, in this work we use cosine distance between vectors to calculate the

Figure 2.3 – One-hot embedding for the sentence “we are not meeting in person.”

We are meeting online.
We are not meeting in person.

We	{ 1 0 0 0 0 0 0 }
are	{ 0 1 0 0 0 0 0 }
not	{ 0 0 0 0 1 0 0 }
meeting	{ 0 0 1 0 0 0 0 }
in	{ 0 0 0 0 0 1 0 }
person	{ 0 0 0 0 0 0 1 }

Source: from the author (2021)

similarity between words. This is used to identify topic-shifts, points where the text changes focus from one subtopic to another, assuming a similarity decrease indicates a change of words and therefore subject.

There are different models for word embedding which are described in further in this Section.

2.4.1.1 Word2Vec

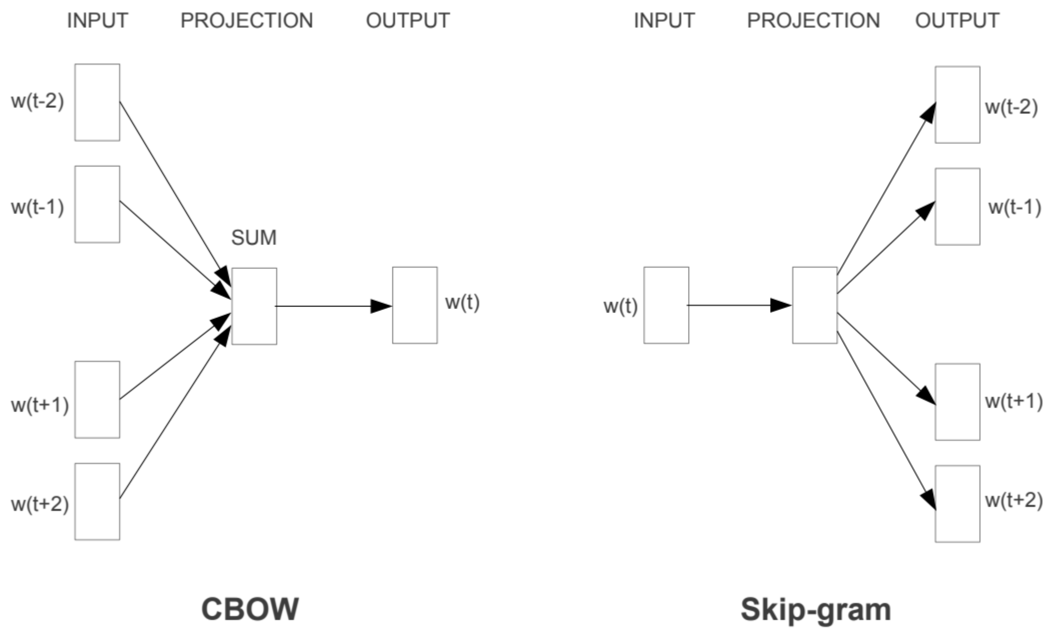
Word2Vec is an algorithm for generating context-free word embeddings developed by Mikolov et al. (2013). The algorithm was built on the idea of the distributional hypothesis, which suggests that words occurring in similar linguistic contexts will also have similar semantic meaning. Word2Vec uses this concept to map words having similar semantic meaning geometrically close to each other in a N-Dimensional vector space.

Word2Vec uses the approach of training a group of shallow, 2-layer neural networks to reconstruct the linguistic context of words. It takes in a large corpus of text as an input and produces a vector space with dimensions in the order of hundreds. Each unique word in the corpus vocabulary is assigned a unique corresponding vector in the space.

Word2Vec is implemented with one of two techniques: Continuous Bag of Words (CBOW) by learning to predict the center word based on the context words, or Skip-Gram by learning to predict given a center word the most likely words in a fixed sized window around it. Figure 2.4 illustrates the process of each technique.

Word2Vec also has its limitations. It depends heavily on the size and quality of the corpus used for training to correctly learn words, and missing words cannot be computed. Although preserving semantic meaning of words, Word2Vec still generates a single vector for each word which can lead to ambiguity.

Figure 2.4 – Word2Vec architecture



Source: Mikolov et al. (2013)

Also, despite using context for learning words, it does not preserve word order. Sentences like *the cat ate a rat* and *the rat ate a cat* would have the same meaning.

2.4.1.2 FastText

FastText is an open-source library developed by the Facebook AI Research lab (BOJANOWSKI et al., 2017). Its focus is to consider the morphology of words when assigning vectors to each word, preventing problems when rare or new words are used. It is based on the skip-gram model from Word2Vec but each word is represented as a bag of character n-grams. The prediction of context words is also implemented differently, defining it as a binary classification task. When learning a word, the words around it are considered positive samples of context words, and random words are selected from the dictionary as negative samples. Then, the algorithm predicts which are context words or not. This allows the processing of unknown words to the model or even mistyped words.

2.4.1.3 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a context-based language modelling technique developed by researchers at Google AI Language (DEVLIN et al., 2019). Its main innovation is training language models bidirectionally, allowing for a deeper sense of language context and flow when compared to previous single-direction language models.

Before BERT, language models would look into a text sequence from from left-to-right or combined left-to-right and right-to-left training, which limited context understanding. As an example, take two sentence: (a) I need to go to the *bank* to make a deposit; (b) I went to the river *bank*. Because the context for the word *bank* is either on the right (a) or left (b), a unidirectional encoder would fail in at least one of the examples. BERT uses both the context on the left and right at the same time, solving this problem.

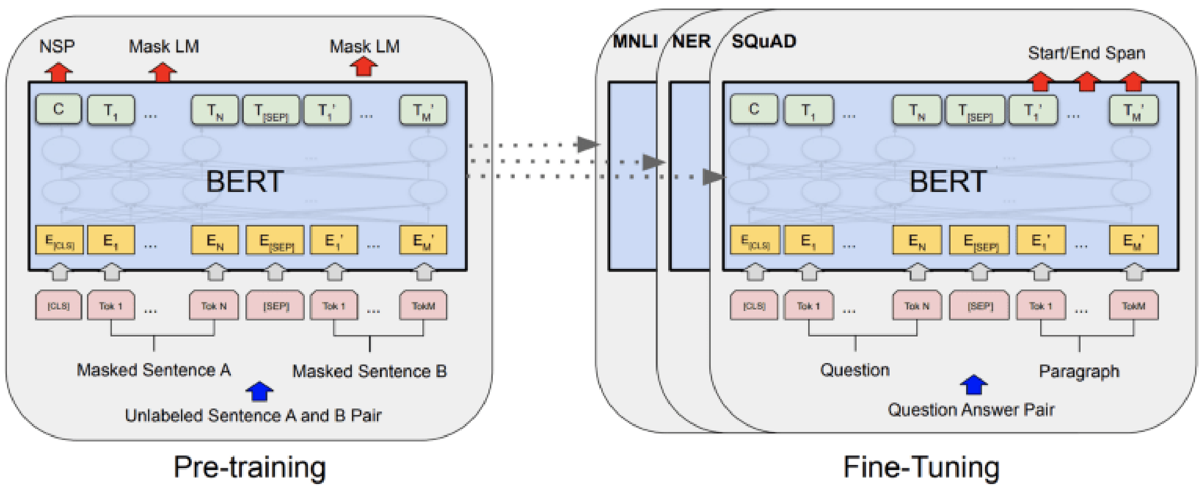
Previous word embeddings such as the aforementioned Word2Vec and FastText create a single vector for each word in their vocabulary, which limits its understanding of ambiguous words. In the previous example, *bank* has two completely distinct meanings but in Word2Vec they both have the same vector. BERT makes use of the context when embedding words which enables it to disambiguate words.

Another challenge in NLP related to BERT is lack of training data. Even though there is an enormous amount of text data available, only few examples of human-labeled training examples are available in task-specific datasets. This is a problem for deep learning researchers because models need huge amounts of data to perform well. BERT helps to fill this information gap by being trained over unannotated text on the web and creating a general purposed model called *pre-training*. When using BERT over a task-specific dataset, it is possible to *fine-tune* it into a task-specific model, still retaining all its pre-trained knowledge.

BERT does not try to predict the next word of a sentence or bag-of-words. Instead, it relies on two strategies:

a) **Masked LM (MLM):** 15% of the words in each sequence are replaced with a [MASK] token. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence. However, this method would create a mismatch between pre-training and fine-tuning the model, since does not appear during fine-tuning. To mitigate this, masked words are not always replaced by [MASK] token. The training data generator chooses 15% of the token positions at random for prediction. If a token is chosen, it is replaced with (1) the [MASK] token 80% of the time (2) a random token 10% of the time (3) the unchanged token 10% of the time;

Figure 2.5 – BERT architecture



Source: Devlin et al. (2019)

b) **Next Sentence Prediction (NSP):** Understanding the relationship between two sentences, which is important in Question Answering and Natural Language Inference tasks which which is not directly captured by language modeling. To understand sentence relationships, the BERT model is trained on a NSP task. The model receives pairs of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original document. During training, 50% of the inputs are a pair in which the second sentence is the subsequent sentence in the original document, while in the other 50% a random sentence from the corpus is chosen as the second sentence. The assumption is that the random sentence will be disconnected from the first sentence.

Figure 2.5 illustrates the pre-training process. A [CLS] token denotes the first sentence and [SEP] separates it to the second sentence to train the NSP task. Both sentences are masked to train the MLM task. The embeddings for each token are fed to the BERT model and the output is both the vector for each word token, the prediction for NSP and predictions for MLM tasks.

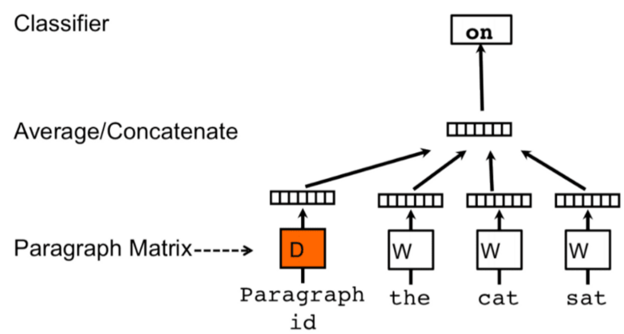
2.4.2 Sentence Embeddings

Sentence embeddings are close to word embeddings in concept. Sentences are embedded in a vector space, retaining some properties from their underlying word embeddings, such as similar sentences beind embedded near each other.

2.4.2.1 Doc2Vec

The Doc2Vec algorithm (LE; MIKOLOV, 2014) is an extension of Word2Vec, which creates embeddings for words. The algorithm follows the assumption that the meaning of a word is given by its neighbouring words. Two variations are presented for Doc2Vec: the Distributed Memory (DM) model and the Distributed Bag-of-Words (DBOW).

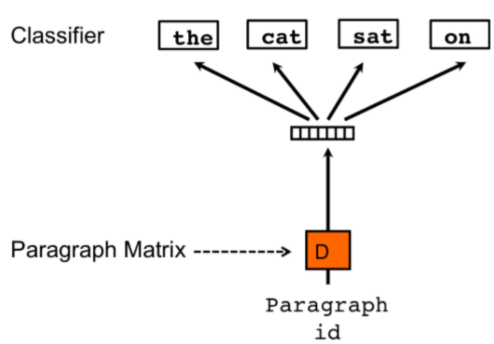
Figure 2.6 – DM model architecture



Source: Le and Mikolov (2014)

In DM, words and sentences of the training corpus are embedded in matrices W and D, respectively. The training is done by sliding a window in the sentence and trying to predict the next word based on the previous, using context and the associated sentence vector. Both the sentence and word vectors are concatenated in a softmax layer to classify the next word. Each step updates the word and sentence vectors. This architecture is depicted in Figure 2.6 The prediction phase is also done with a sliding window over the sentence to predict the next word but only the weights of the sentence vector are updated. After all predictions are computed for a sentence, the resultant sentence vector is the sentence embedding.

Figure 2.7 – DBOW model architecture

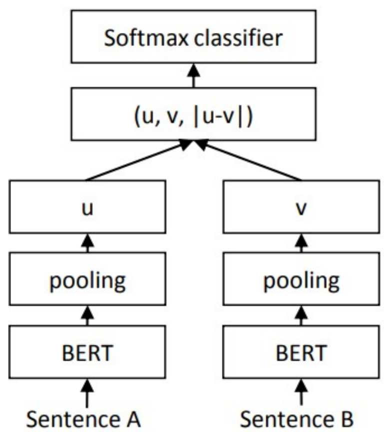


Source: Le and Mikolov (2014)

DBOW ignores word order and has fewer weights. Each sentence in the corpus is converted into a vector. The training is done by selecting a random sentence from the corpus and selecting a random number of words from that sentence. The sentence vector is updated by trying to predict those words based solely on the sentence ID. This architecture is depicted in Figure 2.7. In the prediction phase, a sentence ID is trained with random words from the sentence, but the softmax layer has its weights fixed. The sentence vector is updated in each step, resulting in the embedding for that sentence.

2.4.2.2 Sentence-BERT

Figure 2.8 – SBERT architecture with classification objective function



Source: Reimers and Gurevych (2019)

Sentence-BERT (REIMERS; GUREVYCH, 2019) is a modification of the BERT network that is able to derive semantically meaningful sentence embeddings, enabling BERT to be used in tasks which were not applicable, such as large-scale semantic similarity comparison and information retrieval via semantic search. According to Reimers and Gurevych, finding the similarity of two sentences in a collection of 10.000 would take about 65 hours. By using siamese and triplet networks, Sentence-BERT is able to reduce this time to 5 seconds while maintaining BERT accuracy.

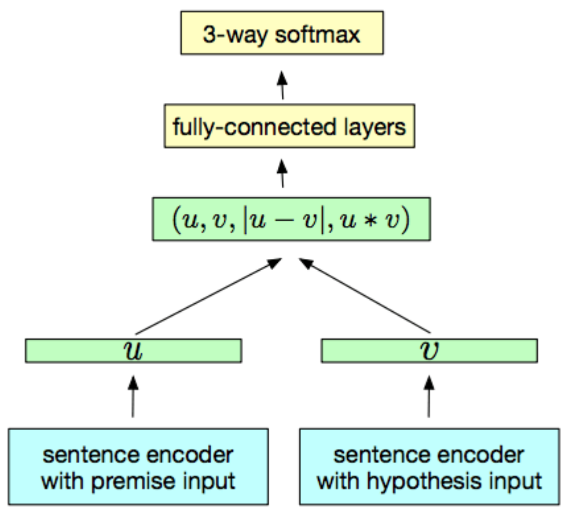
To calculate sentence embeddings, two sentences are computed by BERT, followed by a pooling layer (either max pooling, mean pooling or use the CLS token in BERT output) that generates an embedding for each sentence. siamese and triplet networks are used to update the weights and produce semantically meaningful sentence embeddings which can either have a classification objective function (difference is computed), regression objective function (cosine similarity is computed) or triplet objective function (triplet loss is computed). Figure 2.8 depicts SBERT architecture with classification objective function.

2.4.2.3 InferSent

Researchers at Facebook Research (CONNEAU et al., 2017) created InferSent, a sentence embedding model trained using Natural Language Inference tasks over a BiLSTM network with max pooling.

Natural Language Inference is a task to find a directional relationship between text fragments. For InferSent, the authors used Stanford Natural Language Inference dataset which consists of 570 thousands human generated sentence pairs, each manually labeled with one of three categories: entailment, contradiction and neutral. They believed the semantic nature of this task would make a good candidate to learn sentence representations.

Figure 2.9 – InferSent architecture



Source: Conneau et al. (2017)

Figure 2.9 illustrates the architecture for InferSent. Sentence u (premise) and v (hypothesis) are encoded using BiLSTM. Three matching methods are applied to extract relations between u and v: concatenation of the two representations $(u,\ v)$; element-wise product $u*v$; and absolute element-wise difference $|u-v|$.

The resulting vector captures information from both the premise and hypothesis, and is fed into a 3-class classifier consisting of multiple fully connected layers followed by a softmax layer.

2.5 Related works in topic labeling

Topic labeling aims to assign a descriptive phrase to represent topics discovered in a document. This task is often divided into two steps: candidate label generation and candidate label ranking. The first deals with how candidate labels are generated, common approaches are using n-gram co-occurrence (MEI; SHEN; ZHAI, 2007; MAO et al., 2012), top-n words (LAU et al., 2010), and using external resources such as Wikipedia (LAU et al., 2011; BHATIA; LAU; BALDWIN, 2016), Google Directory (MAGATTI et al., 2009) or Twitter (BASAVE; HE; XU, 2014). After generating candidates, the second step is to rank labels based on semantic relevance scoring methods.

Mei, Shen and Zhai (2007) defined the topic labeling problem as "Given a set of latent topics extracted from a text collection in the form of multinominal distributions, our goal is, informally, to generate understandable semantic labels for each topic". Their candidate label generation is based on two techniques: parsing the text to extract frequent chunks or phrases, and n-gram co-occurrence. The label ranking is based on the probability that a word is important for its topic, labels with more important words are ranked higher.

Instead of using co-occurrence, Magatti et al. (2009) proposed using Google Directory hierarchy to generate candidate labels. Their assumptions was that the world is described by a set of concepts (or topics) and could be inserted into a light ontology to generate candidate labels with ontological alignment of topics by building a Topics Tree using Google Directory hierarchy. The label ranking process is done by using similarity measures, namely: cosine, overlap, mutual, Jaccard and Tanimoto similarity. Basically, the candidate label with highest similarity over all alternatives is selected as the topic label.

Lau et al. (2010) proposed to approach topic labelling via best term selection, more specifically selecting one of the top-10 topic terms to label the overall topic. While it is often possible to label topics with topic terms, there are also often cases where topic terms are not appropriate as labels. Words outside the top-10 can by chance be better suited to label the topic, and it is often better to label a topic using multiword terms.

Lau et al. (2011) is a natural progression from their previous work. They proposed to use Wikipedia articles as topic labels, the problem then shifts to finding relevant Wikipedia articles to be used as labels. The top-10 words from each topic are used to query Wikipedia and acquire candidate articles. In the ranking process, they used several lexical association measures: point-wise mutual information, Student's t-test, Dice's coefficient, Pearson's χ^2 test, the log likelihood ratio (PECINA, 2009), and the conditional probability and reverse conditional probability measures based on Lau et al. (2010).

Building on top of the work of Lau et al. (2011) and bringing neural embeddings to topic labelling algorithms, Bhatia, Lau and Baldwin (2016) developed NETL (neural embedding topic labelling) [9]. It labels documents using titles from Wikipedia articles and applying Word2Vec for label candidate generation. Instead of using the native search API and a site-restricted Google search to retrieve label candidates, Bhatia, Lau and Baldwin (2016) improved this method by training word embeddings models on top of a Wikipedia dump, removing the need for external resources, and improving the system's accuracy.

The top-10 terms for each topic are used to measure the relevance of each title embedding based on the cosine similarity with each of the word embeddings. The Doc2Vec relevance(rel_{d2v}) and Word2Vec relevance (rel_{w2v}) of a title a and a topic T is given as follows:

$$rel_{d2v}(a,T) = \frac{1}{|T|}\sum_{v \in T} cos\left(E^{d}_{d2v}(a), E^{w}_{d2v}(v)\right) \tag{2.2}$$

$$rel_{w2v}(a,T) = \frac{1}{|T|}\sum_{v \in T} cos\left(E^{w}_{w2v}(a), E^{w}_{w2v}(v)\right) \tag{2.3}$$

where $E^{d}_{d2v}(a)$ is the document embedding of the title a generated by Doc2Vec; $E^{w}_{d2v}(v)$ is the word embedding of word y generated by Doc2Vec; $E^{w}_{w2v}(v)$ is the word embedding of word z generated by Word2Vec; $v \in T$ is a topic term; $|T|$ is the number of topic terms; and $cos(\overrightarrow{x}, \overrightarrow{y})$ is the cosine similarity function.

Both Word2Vec and Doc2Vec are used to generate title embeddings because, as Bhatia, Lau and Baldwin (2016) observed, Doc2Vec tends to favours fine-grained concepts, while Word2Vec favours generic or abstract labels. This difference is a consequence of the characteristics of each model. While Doc2Vec uses the words from the title to determine the title embeddings, Word2Vec uses the neighbouring words of the title token in the text directly, oblivious to the composition of words in the title.

The following formula is used to combine the strengths of Doc2Vec and Word2Vec by summing the relevance scores using top-100 candidates from each model:

$$rel_{d2v+w2v}(a,T) = rel_{d2v}(a,T) + rel_{w2v}(a,T) \tag{2.4}$$

The candidate labels are then ranked based on four features: 1) Letter trigram, based on the finding of Kou, Li and Baldwin (2015), which is an effective way of ranking topic labels; 2) Page Rank (PAGE et al., 1999), which is used to favour labels that represent more "core" concepts in Wikipedia; 3) The number

[9] Available at <https://github.com/sb1992/NETL-Automatic-Topic-Labelling->

of words in the candidate label; 4) Topic overlap is the relative number of terms in the candidate label and the top-10 topic terms. The last two features were proposed by Lau et al. (2011).

This Section presented concepts used through this project, describing strategies used by visually impaired people when navigating the Web, topic segmentation algorithms, and topic labelling algorithms. The next Section presents a systematic mapping conducted to gather related work on automatic web content processing strategies for visually impaired users.

2.6 Related Work on Automatic Web Content Processing Strategies for Visually-Impaired Users

This section presents a literature review covering studies related to the problem addressed by this project. The analysis results from a systematic mapping to have a better grasp of previous studies that aimed to help screen reader users in information seeking tasks –the methodology is available in Appendix A. Our queries in 4 databases (Scopus, Engineering Village, ACM Digital Library and Web of Science) returned 235 papers, which were filtered by applying inclusion criteria. The final result was 16 works shown in Table 2.3.

Throughout the study, the contributions reported in these papers were used as a basis for design decisions and comparisons. The reviewed studies were separated into different categories based on the approach used (descriptions of each category are in the appendix). The categories were Text Reduction, Content Filtering, Navigation, Concurrent Speech, Recommendation Systems and Auditory Overview.

2.6.1 Text Reduction

One of the most common techniques to aid visually-impaired people when reading long texts on the Web with screen readers is reducing the length of a given text. In the results of this systematic mapping, nine out of sixteen papers addressed this method. Thus, this category has the widest variety of algorithms, with summarization being the most common approach.

Ahmed et al. (2012) aimed at achieving a non-visual skimming method. For this, the authors proposed automatically generating a text summary, selecting words from each sentence of a given text independently. Each sentence was analyzed separately without taking into account the context or the rest of the text. Summarizing each sentence by itself was necessary because, according to the authors, texts on the web tend to be shorter, usually no more than five paragraphs. Therefore, it would not be appropriate to use an approach that relies on statistical parameters because it is less efficient in short texts than in long texts.

Table 2.3 – Selected Articles

Reference	Year	Research Type	Approach Type
YARRINGTON; MCCOY	2009	Philosophical Paper	Content Filtering
LI et al.	2010	Evaluation Research, Solution Proposal	Text Reduction, Recommendation
AHMED et al.	2012	Evaluation Research, Solution Proposal	Text Reduction
GHAHARI et al.	2012	Evaluation Research, Solution Proposal	Navigation
AHMED et al.	2013	Evaluation Research	Text Reduction
YE; LI; LI	2014	Evaluation Research, Solution Proposal	Text Reduction, Recommendation
GADDE; BOLCHINI	2014	Validation Research, Solution Proposal	Content Filtering
GUERREIRO; GONÇALVES	2014	Evaluation Research	Concurrent Speech
YE; LI; LI	2014	Evaluation Research, Solution Proposal	Text Reduction, Recommendation
MANISHINA et al.	2016	Solution Proposal	Text Reduction, Concurrent Speech
GUERREIRO; GONÇALVES	2016	Evaluation Research	Concurrent Speech
GUERREIRO; GONÇALVES	2016	Philosophical Paper	Concurrent Speech
TRIPPAS	2016	Philosophical Paper	Text Reduction, Navigation
WANG; REDMILES	2017	Solution Proposal	Auditory Overview
ALVES; CARDOSO; FREIRE	2018	Validation Research, Solution Proposal	Text Reduction
GIRAUD; THÉROUANNE; STEINER	2018	Evaluation Research, Solution Proposal	Content Filtering

Source: from the author (2021)

The automatic summarization was developed using a classifier algorithm that was trained using machine learning techniques. The Stanford Parser (KLEIN; MANNING, 2003) generates characteristic vectors for each word of each sentence of the text. The Stanford Parser is a probabilistic natural language parser. It can analyze sentences' grammatical and syntactic structure and extract relations and POS (parts-of-speech) tagging of the words. Using the grammatical relations generated by Stanford Parser, it is possible to construct a directed graph. In this graph, nodes are words, and directed edges are the relations from the governor word to the dependent. Based on this, a tree is generated based on the grammatical relationships between words. The classifier selects the most relevant words in the tree. A tree connection algorithm is used to select supplementary words to keep the text flowing to a minimum. This process ensures all selected nodes have grammatical relations between them and increase comprehension of the skimmed sentence.

The authors conducted an evaluation study with 23 blind participants who read a human-made summary and the automatic summary to evaluate the tool. Participants were required to complete two tasks composed of sets of questions. The first was a comprehension test, where the participant should read the whole text before answering comprehension questions. The second was an information-seeking task, and users had to search for the answers to the questions in the text. The tests showed that the improvement in performance and time spent on each task was similar between the two summaries. When asked, participants were unable to identify which summaries were made automatically.

Despite the positive result, the tests were done with texts that were previously summarized. To effectively evaluate the developed tool, it is essential to perform tests in a real scenario, some factors that affect user experience are left out in with preprocessed text, such as the processing time and responsiveness. The summarization method can be evaluated for real-time effectiveness during site navigation and to measure the time required to generate the summary.

In a follow-up study, Ahmed et al. (2013) applied the algorithm mentioned above for touch-screen devices with a notable difference. A variable-sized summary generation algorithm was added. This algorithm associated a confidence score (between 0 and 1) with each word to be included in the summary. A slider was added to enable the user to control the text compression. Based on the value of this scale, the words were selected for the resulting summary (*e. g.* if the slider is at 50% compression, only words with confidence above 0.5 were selected). Their user study with 15 participants also showed promising results, reducing the time required to answer more than half of the read texts.

To read news with smartphones, Ye, Li and Li (YE; LI; LI, 2014b; YE; LI; LI, 2014a) used a multi-document summary to ease information retrieval for screen reader users by combining similar news articles into one summarized text. First, relevant news articles are filtered out using an adjusted DSYNT tree (Deep-Syntactic tree) (BARZILAY; MCKEOWN; ELHADAD, 1999). After that, they employ SumBasic (NENKOVA; VANDERWENDE, 2005), a multi-document summarization algorithm, to generate a summary for these filtered relevant news articles.

Manishina et al. (2016) used TF-IDF and HTML parameters that reflect the visual properties of each word in the document, such as colour and size. They were used to identify each word's importance in order to select key terms from the text. The terms extracted are used to segment the web page into zones to be read in a concurrent speech by the screen reader to represent the dense visual stimulus of the web page. This approach is further explained in Section 2.6.4.

Trippas (2016) investigated the use of a summary for queries, observing the user preference between short or long summaries for search results. The data showed that users preferred longer and informative summaries for text presentation, but shortened summaries were preferred for audio queries.

Alves, Cardoso and Freire (2018) used two superficial methods, which use only statistical data and few linguistic-computational resources to summarize texts on the Web. These methods were the bushy path and keyword extraction methods, both with a low computational cost. The excerpts of the generated text in that paper were used as internal links inserted into the web page. Each link pointed to the corresponding part of the text from where it was generated.

2.6.2 Content Filtering

Content filtering includes algorithms that filter content from a web page to be read by the screen reader or sections of web pages. The common idea in these papers is that reducing the amount of content available for the screen reader user can lower the cognitive load since the user is exposed to less content, simulating what sighted users do while glancing at a page.

To facilitate navigation between sections of a web page, Gadde and Bolchini (2014) conducted a study aimed at identifying common navigation problems encountered by blind users, identifying which sections are most important at each purchase step on the Amazon website, and developed a tool that can suggest user sections based on an importance ranking. A significant result of this study was the identification of navigation problems, which, in summary, are described as follows:

a) Even simple pages obstruct navigation due to their structure;
b) Rigid access to different page types may make navigation difficult;
c) Fear of missing crucial information forces unnecessary cognitive and physical load;
d) Little tolerance of navigation mistakes demands increased attention;
e) Blurred sections and overloaded pages can obstruct access to important sections;
f) 'Skip to content'[10] features may not be optimal;
g) Difficulty to find previous information forces the user to remember details and overloads short-term memory.

[10] Shortcut often located on top of a web page to enable screen reader users to quickly find the main content of the page

Gadde and Bolchini (2014) used a crowd-sourcing approach to rank page sections according to their importance. This study analysed the Amazon online store, and the importance of each section was based on its weight towards the buying decision. Based on this ranking system, the user could use the proposed tool to jump directly to specific sections according to their importance, making navigation more straightforward and practical.

The user tests in that study relied solely on one experienced user, serving more as a preliminary study than actually a user evaluation. Better validation would require more users and a variety of expertise in the use of screen readers. Another shortcoming of the study is the reliance on the previous ranking of page sections. The authors cite future work as using HTML tags to provide website developers with tools to inform the software of the importance of each section. However, that would depend on the developer's knowledge.

In order to aid the navigation of screen reader users, Giraud, Thérouanne and Steiner (2018) proposed a web content filtering approach. A tool was developed with features to filter redundant and irrelevant information for the user's tasks. Redundant information is presented more than once, for instance, a website logo or a menu repeated on different pages, and the irrelevant information was information that added nothing to the user tasks. In that approach, when a user enters a site, he/she hears all the relevant information contained therein. However, in the following actions during the same session (such as when the user goes to another site page), redundant information is ignored.

To test the efficiency of the developed tool, three user tests were conducted, with 76 users in total. The scenario was simulated in the first two tests, while in the third test, it was in a real scenario. In the first test, 25 participants used the tool for predefined tasks. They then answered a questionnaire that measures cognitive load, NASA-RTLX (NASA Raw Task Load Index)[11] (HART; STAVELAND, 1988). In the second test, the procedure was the same, but a secondary task, which consisted of pressing "F" each time a specific sound was played, was added to assess the cognitive load instead of the NASA-RTLX. Finally, in the third test, the procedure was similar to the first one. However, in the end, the SUS (System Usability Scale) (BROOKE et al., 1996) questionnaire[12] was also applied to evaluate the usability of the system.

[11] A questionnaire composed of six 7-point scales to measure different factors of cognitive load: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, Frustration.

[12] A 10 item questionnaire with five response options ranging from "Completely Agree" to "Completely Disagree", based on a Likert scale.

As a result, a significant decrease in cognitive load could be observed between filtered and unfiltered scenarios. In the filtered scenarios, the cognitive load and the average time to finish the tasks were smaller. Satisfaction was significantly higher, and the error percentage was not significantly different.

Yarrington and McCoy (2009) also proposed a content filtering approach. In that paper, the focus was on answering questions given by the user by filtering areas of the text that could answer them. To do the filtering, the keywords of the document were weighted using a variant of TF-IDF. In this version, each word was weighted by the inverse of the number of paragraphs where they occur in the document. The idea is that if a word appears in every paragraph, it is likely to be related to the overall topic of the document instead of specific areas of interest. They also used WordNet (MILLER, 1998), an English lexical database, to include synonyms, hyponyms and hypernyms of words in questions to increase the chance of finding the correct answer within the text but both results, with or without WordNet, were limited.

2.6.3 Navigation

This category included studies that aimed to help visually impaired users by giving a better navigation experience. Trippas (2016) approached this by using a conversational search. She used search engines to test their usability in a speech-only interaction and concluded that further work needs to be done to develop techniques that better suit a spoken interaction with query suggestions and better results listings.

One problem commonly faced by visually impaired people is retrieving information that had already been encountered. Sighted people only need to remember the position where the information was located, but visually impaired people need to hear to most of the page content again. One way to reduce this problem that has already been implemented is by improving back navigation shortcuts. To assess this issue, topic-based back navigation was developed by Ghahari et al. (2012), where a back shortcut is used to go back to previous pages based on related topics; and list-based back navigation, where users can quickly go back to a page with a list of links (such as a page listing professors of a university). A user study was conducted with ten screen reader users, and it showed promising results. Time-on-task was reduced by almost half, the number of backtracked pages and keystrokes for backtracking was reduced by more than half, and the navigation experience benefited from lower cognitive efforts.

2.6.4 Concurrent Speech

Looking for an alternative way of conveying information contained in texts to screen reader users, Guerreiro and Gonçalves (2014) proposed the use of concurrent speech. In this method, two audio tracks are played simultaneously so that the user can find information faster. The study takes advantage of the *Cocktail Party Effect* (CHERRY, 1953), which states that humans can focus their attention on one sound source among several others, and also detect the content of interest in background voices. The authors evaluated the perception of concurrent speech by blind people, finding that two and three talkers can identify more information in less time.

In 2015, these authors evaluated the impact of speeding up concurrent speeches in comparison to one-voice speech and found that two and three-voices with speech rates slightly faster than the default can enable a significantly faster scanning for information still maintaining overall comprehension (GUERREIRO; GONÇALVES, 2016). In the same scenario, to keep up with concurrent speech completion times, one-voice would require faster reading, causing a more significant loss in comprehension and performance.

Manishina et al. (2016) developed a strategy to enable fast access to web page content in non-visual situations taking into account page layout and typographic clues. As explained in Section 2.6.1, the content of a web page is segmented into different zones. These sections are read using concurrent speech, using many *Cocktail Party Effect* metaphors. The first metaphor is the repetition frequency: the larger the group, the more often terms about a topic will emerge. Based on this, the larger the zones, the more frequent the key term is read. The second metaphor was volume. The volume of speech was determined by the contrast between the text and background of each zone, and the frequency of the key term. The third was spatialization: the voice and 2D spatialization of vocalized key terms changed according to zone coordinates.

2.6.5 Recommendation Systems

Considering the overflow of information present on the Internet, finding relevant content may be a time-consuming and challenging task. With this in mind, (YE; LI; LI, 2014b; YE; LI; LI, 2014a) developed ways to recommend news based on the user navigation history with an ontology-based approach. First, TF-IDF is used to select keywords to represent the document being read. Second, an ontology tree is used to calculate word similarity between documents, giving more weight to word present in the titles. The historical readings from the user are also taken into account when recommending new documents. In their use study, 113 visually impaired participants were invited, and results showed high levels of satisfaction with the recommendations and relevance of recommended articles.

Another recommendation system was implemented by Li et al. (2010), where a system denominated RAIN recommends sections of a page that might be relevant but ignored by the user. It takes advantage of the fact that users with similar intentions might navigate through similar paths. The RAIN system was tested with five blind users and four tasks. The results show that when using the recommendation system, users could complete tasks twice as fast.

2.6.6 Auditory Overview

To simulate the action of glancing a web page to get an overview, a browser extension was developed by Wang and Redmiles (2017) to summarize the accessibility features found in a page in a shot and dynamically generated soundtrack. This way, users can have a quick overview of useful elements present on a web page. In this implementation, the presence of elements such as headers and links produce sounds on the soundtrack, for instance, birds and barking sounds.

This systematic mapping gave us an overview of the different strategies approached by other researchers. We can observe a predominance of text reduction techniques over the alternatives and has been explored in different ways. However, those solutions still have lower availability on popular screen readers. This also fitted best our research interest, and we expect that, after implementing this strategy into a screen reader that is widely used by the Brazilian public, we can help users in their daily lives. Recommendation systems, content filtering and navigation strategies were also relevant strategies. However, some studies showed they could be more challenging to use, given the lack of standardization of the Web and how often websites do not respect accessibility guidelines. Two other newer approaches were also found, auditory overview and concurrent speech, both using the audio properties in screen readers. As those strategies had more limited evidence on the use, we believed they would not be viable to investigate in the Brazilian scenario with Brazilian Portuguese at the moment.

This chapter presented a background of essential concepts used throughout this work, such as accessibility, summarization, and topicalization techniques. The chapter also presented the results of a systematic mapping with research aimed to help screen users seeking information when searching the Internet or reading longer texts. The next chapter provides information about the methods used to develop the proposed screen reader add-on and test it with visually-impaired users.

3 IMPLEMENTATION OF THE HEADER GENERATING ALGORITHM

In this study we aimed at automatically generating headers inside long texts in order to facilitate information seeking tasks for visually impaired users. By inserting headers inside long texts, users would be able to infer the content of a section by reading the header, or use it to mentally map where each information is inside the text, aiding information retrieval.

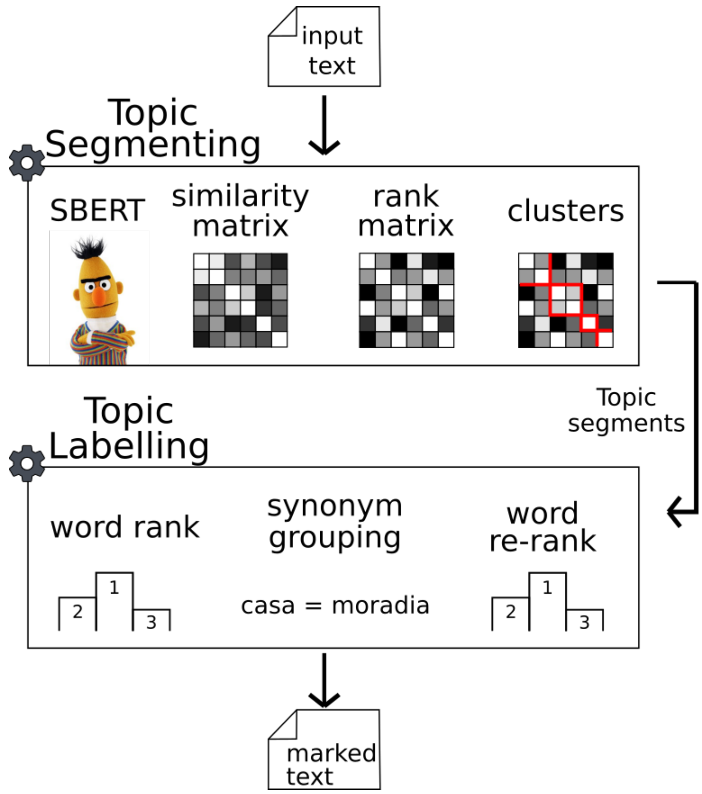
The task of automatically generating headers can be divided in two algorithms: one for segmenting a text into multiple topics, the other to assign a label that comprises its general subject. Topic segments are sequences of subtopics that are assumed to occur during the discussion of a main topic. In this work we identify topics by analyzing a sentence-level discourse structure based on **topic shift** to subdivide expository texts into passages with multiple sentences (i.e. the unit used to segment topics are sentences). This chapter describes the topic segmenting algorithm and topic labelling algorithm developed for this purpose. Since NVDA is the most popular screen reader in Brazil, and it is implemented in Python, we decided to implement our algorithms using the same programming language for a easier integration with the screen reader.

Our algorithm framework is illustrated in Figure 3.1. The topic segmenting algorithm uses a BERT model to compare each and every sentence of the input document, creating a similarity matrix to which a filter is applied to create a rank matrix. BERT was chosen over other options to observe how this new model would perform in this classic algorithm. Using this matrix, the algorithms infers topic shift points in the document, subdividing the text into clusters of contiguous sentences, which are called topic segments. The topic labelling algorithm processes each topic segment separately by ranking words and grouping its synonyms (avoiding redundant words) to automatically generate a header. This process results in a text segmented into text blocks identified by labels (or headers) which can easily be adjusted other purposes (such as HTML injection or tags for screen readers). All process is explained in more detail in the following Sections.

3.1 Automatic Topic Segmentation Algorithm

The segmentation algorithm developed in this study was based on ToSe-Word2vec (NAILI; CHAIBI; Ben Ghezala, 2017), which combines word embeddings with C99 (CHOI, 2000). In this study we decided to use BERT, a state-of-the-art model, to calculate similarities. BERT was chosen over other options to observe how this new model would perform in this classic algorithm. The algorithm receives a list of sentences parsed with NLTK and outputs a marked file with sentences divided by topic. It is worth noting that in the

Figure 3.1 – Algorithm framework



Source: from the author (2021)

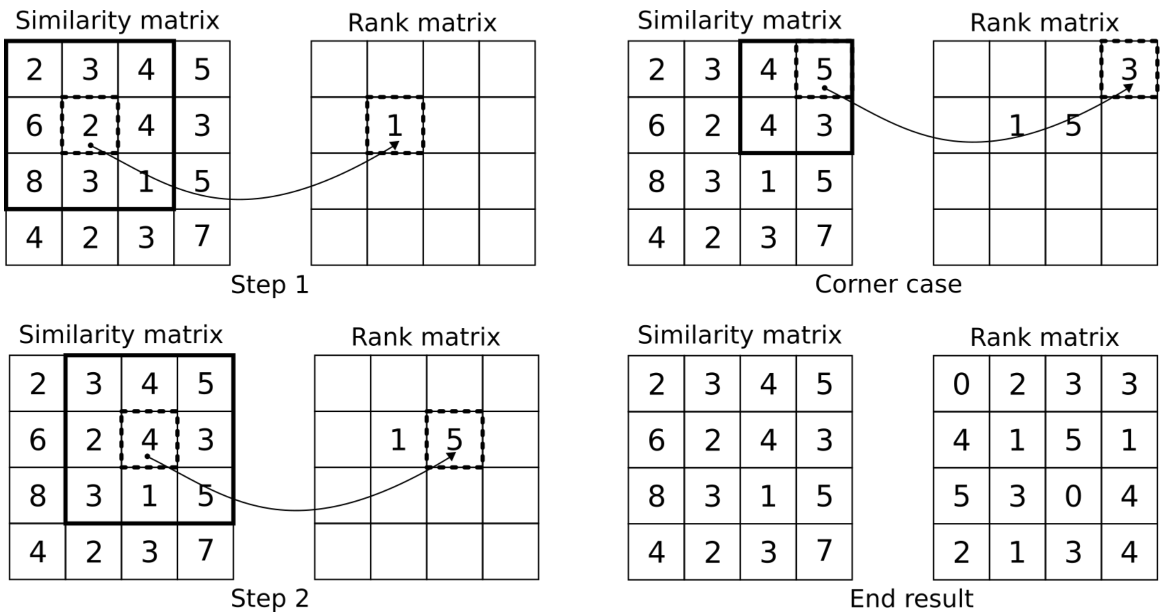
parsing process, we lose any hint on where a paragraph ends, thus paragraphs can be split in the middle if the algorithms finds a topic boundary between two of its sentences.

For a text of n sentences, a similarity matrix $n \times n$ is created using SBERT[1] (REIMERS; GUREVYCH, 2019), a python library for sentence embeddings. We used a multilingual BERT model to enable comparing the segmentation results with previous research, we opted for using distiluse-base-multilingual-cased-v1[2] (REIMERS; GUREVYCH, 2020), which was readily available for usage through the SBERT library. The similarity is calculated using cosine distance over the embedding of two sentences being compared. Because we are comparing small text fragments (sentences), the absolute similarity value is unreliable. According to Choi (2000), a small change on the numerator can cause a disproportionate increase in similarity, unless the denominator (number of tokens) is large enough. Because of that, the similarity value cannot be used to statically compare how much a sentence is more similar than the other. It is only possible to estimate the order of similarity between sentences, *e.g.* sentence a is more similar to b than c. To mitigate this it is

[1] Available at <https://www.sbert.net/>
[2] Available at <https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1>

necessary apply a filter using an 11×11 rank mask to create a rank matrix, where the rank for a cell is the number of its neighbouring cells with lower similarity value.

Figure 3.2 – Example of rank matrix



Source: from the author (2021)

$$r = \frac{\text{number of cells with a lower value}}{\text{number of examined cells}} \tag{3.1}$$

Figure 3.2 illustrates the rank masking process in a 4×4 matrix and an 3×3 rank mask. In Step 1 the dotted cell is being calculated and it is compared to its neighboring cells. Inside the 3×3 square, just one cell has a smaller value, resulting in its rank score of 1. The process continues for each cell to complete the rank matrix. To avoid normalisation problems (e.g. when calculating a cell in a corner), the rank is expressed by ratio r (equation 3.1).

The last step is to detect topic boundaries by clustering sentences using Reynar's maximisation algorithm (REYNAR, 1994). A text segment composed by two sentences i and j is represented by a square region along the diagonal of the rank matrix. Let $s_{i,j}$ denote the sum of the rank values in a segment and $a_{i,j} = (j - i + 1)^2$ be the inside area. $B = \{b_1, ..., b_m\}$ is a list of m coherent text segments. s_k and a_k refers to the sum of rank and area of segment k in B. D is the inside density of B (equation 3.2). Choi (2000) recommends $c = 1.2$, but for our case, this value was generating segments too short. After running tests to generate larger text segments by gradually increasing the value of c, we observed $c = 2$ achieved the desired result while still improving its window difference score. For these tests we used a selection of 4 large Wikipedia articles while pinpointing the c value, and then tested it on CSTNews (Portuguese) and Choi dataset (English) to confirm if metrics actually improved. Table 3.1 shows the difference in Window Difference and P_k scores for $c = 1.2$ and $c = 2$.

$$D = \frac{\sum_{k=1}^{m} s_k}{\sum_{k=1}^{m} a_k} \tag{3.2}$$

Table 3.1 – Improvement in scores with c value increase

	Wikipedia articles		Choi		CSTNews	
c	WD(%)	Pk(%)	WD(%)	Pk(%)	WD(%)	Pk(%)
1.2	58,80	46,80	18,00	16,00	39,50	36,20
2.0	43	39,70	16,30	15,70	34,20	32,20

Source: from the author (2021)

At the start of the process, the entire document is considered a single coherent text segment B. Each step of the process splits one of the segments in B. The split point is a potential boundary which maximizes D. The number of segments to generate, m, is determined automatically by applying the threshold $\mu + c \times \sqrt{v}$, where μ and v are the mean and variance of $\delta D^n, n \in \{2, ..., b+1\}$, and c is the variable used to control the threshold sensitivity. δD^n is the gradient $\delta D^n = D^n - D^{n-1}$ and D^n is the density of n segments.

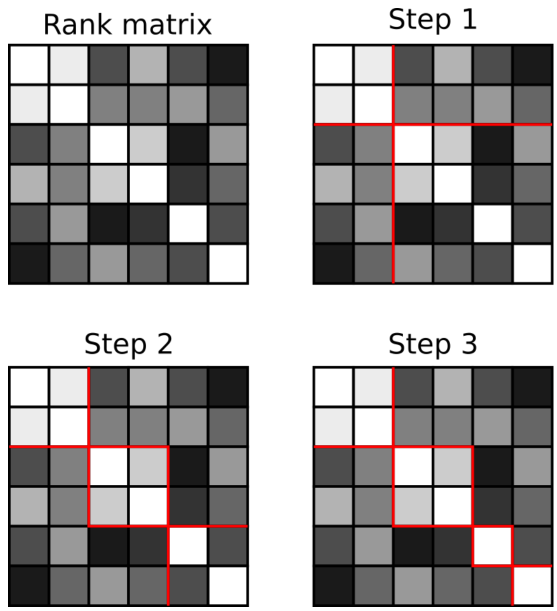
Figure 3.3 illustrates this process. At the start, the entire rank matrix is considered as a coherent text segment. High similarity is represented as clearer colors. The algorithm detects a boundary between the second and third sentences, where it is observable that colors became darker, meaning low similarity between sentences. Since only this boundary was detect in Step 1, the third sentence onwards is considered as a coherent segment. The same happens in Step 2, forming another cluster with sentences 3 and 4. It is worth noting that, even though we can see that there is gray values inside this second cluster, the difference is insufficient to pass the threshold, so it is still considered coherent. Step 3 finishes the process detecting the last boundary, resulting in 4 text segments detected for this example.

3.2 Automatic Topic Labelling Algorithm

Our labelling algorithm assigns keywords for each topic segment individually. Words are ranked based on frequency, the idea is that if a word is repeated in a text fragment, it might reflect its subject. We first remove stopwords [3] to ensure the label is consisted of only meaningful words.

[3] Stopwords are words that are very common in a language but adds little value to the task being done. There is no universal stopword list.

Figure 3.3 – Example of the clustering algorithm



Source: from the author (2021)

We started our stopword list with the default from NLTK, and added other words that added little value for labels. We also filtered words based on its part-of-speech tag[4] based on Ahmed et al. (2012) findings that nouns, verbs and adjectives were the words that most contributed semantically for a sentence. Since NLTK lacks a part-of-speech tagger for Portuguese, we used UDPipe (STRAKA, 2018), a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files.

Using just the word frequency caused to problems: words with similar meaning in the labels, and important words not appearing in the label due to the use of synonyms to avoid repetition. The first issue makes words that add little significance for the label being chosen, since they have similar meaning. The second is caused by the fact that, in well written text, writers tend to use synonyms to avoid word repetition, which would affect word count making a relevant word be lower ranked than it should. To avoid both problems, we group each word with its synonym sets (synset). For each word in the text, a synset is created using a combination of all its synonyms in the wordnet available in NLTK and TEP 2.0 (MAZIERO et al., 2008). We used two sources for this step so one would complement the other when it lacks a rare word, or if only a few synonyms are included. In some cases such as "computador", TEP2 lacked synonyms while wordnet had 6 words. In other cases such as "celular", wordnet only had synonyms for mobile phones

[4] Part-of-speech are category of words with similar grammatical properties. For Portuguese they are: noun, adjective, verb, pronoun, adverb, numeral, article, preposition and interjection.

('telefone', 'telemóvel') while TEP2 had synonyms from biology field ('celuloso', 'celulado'). The idea of this strategy is to cover a wider range of synonyms. We did not apply any word-sense disambiguation.

When processing a topic segment, if a word already exists in a synset, its count is added to it, meaning only the first word that appeared in the text from the set is used for the label. As an example, when "menina" appears in a text, its synset is created as 'menina': menina, garota, guria. If "garota" appears thereafter, its count is added to "menina" because it is already in its synset.

In order to search for a word in wordnet or TEP2, it is necessary to use its dictionary form, also called lemma. Inflected forms of a word, such as verb tense or plural, are not included in these word collections. Inflected words need to be lemmatized, that is, transformed back to its dictionary form. As an example, the word *"fiz"* is in the past tense and its lemma is the verb *"fazer"*. The lemmatization process was done with UDPipe, since NLTK does have a lemmatizer but it is unavailable for Portuguese. We also experimented with Spacy, an open-source library for Natural Language Processing, but the lemmas created were not as consistent as UDPipe. Words were also lower cased in order to search them in TEP2 and wordnet.

The labels generated by our algorithm have a varying number of keywords because our topic segmenting algorithm creates topics with different sizes. A topic might be too small for the number of keywords used to label it, making irrelevant words appear whilst a topic too large would not have enough keywords to comprise all its subjects. A varying number of keywords would enable labels to be suited for both cases. After testing the label size ad-hoc over CSTNews[5], we decided that topic labels should vary from 3 to 5 keywords based on word count for each text segment. This assures that larger segments have more words to describe them and smaller segments will not add irrelevant words due to small word count.

Figure 3.4 shows an example snippet of a text about Queen Elizabeth II taken from Wikipedia with topic separation and headings generated by the algorithm. This example is composed of 14 sentences and three paragraphs. The first topic (Figure 3.4a) has more words in the header because it is long, we believe this helps the user to have a better understanding of its content. The second 3.4b is too short to rank 5 keywords, thus only 3 are used. In fact, after the third word used in the heading, the next would simply follow the order they appeared in the topic, since no other repetitions occurred.

The first segment is composed of 10 sentences and almost three paragraphs from the original. The last word of each paragraph from the original text are underlined, it is possible to notice the two first paragraphs and half of the third were included in the first segment. It describes who Isabel II is, which countries she is queen of, and a bit about her early life before being crowned.

[5] A corpus of journalistic texts in Brazilian Portuguese available in: https://sites.icmc.usp.br/taspardo/sucinto/cstnews.html

Figure 3.4 – Label examples

(a) ISABEL INDEPENDENTE ESTADO RAINHA REINO

Isabel II ou **Elizabeth** II (Londres, 21 de abril de 1926) é a atual **Rainha** do **Reino** Unido e de quinze outros **estados independentes** conhecidos como **Reinos** da Comunidade de Nações, além de chefe da Commonwealth formada por 53 **estados**. Ela é a primeira monarca feminina **soberana** da Casa de Windsor e a Governadora Suprema da Igreja da Inglaterra. Em algum de seus outros **Estados soberanos**, ela possui o título de Defensora da Fé. **Isabel** se tornou a Chefe da Comunidade Britânica quando assumiu o trono em 6 de fevereiro de 1952, após a morte de seu pai, o rei Jorge VI. **Isabel** é a **rainha** de quatro países **independentes**: **Reino** Unido, Canadá, Austrália e Nova Zelândia. Entre 1956 e 1992 o número de países do **reino** britânico variou já que certos territórios ganharam sua **independência** e outros tornaram-se repúblicas. Atualmente, além dos quatro primeiros **estados** mencionados, **Isabel** é **rainha** da Jamaica, Barbados, Bahamas, Granada, Papua-Nova Guiné, Ilhas Salomão, Tuvalu, Santa Lúcia, São Vicente e Granadinas, Belize, Antígua e Barbuda e São Cristóvão e Nevis. **Isabel** nasceu em Londres e foi educada particularmente em casa. Após a Crise da abdicação de Eduardo VIII, tio de **Elizabeth**, houve a ascensão de George VI ao trono, o que fez de sua primogênita a herdeira presuntiva da coroa. **Isabel** passou a assumir deveres públicos durante a Segunda Guerra Mundial, em que ela serviu no Serviço Territorial Auxiliar.

(b) ANO CASAR PRÍNCIPE

Ela se **casou** com o **príncipe** Filipe da Grécia e Dinamarca em 1947, com quem teve quatro filhos: Carlos, Ana, André e Eduardo. Seu pai morreu em fevereiro de 1952 e Isabel ascendeu ao trono aos 25 **anos**. Sua coroação ocorreu no **ano** seguinte e foi a primeira a ser televisionada.

Source: from the author (2020)

The second paragraph has just 3 sentences. Interesting enough, the second segment ends in the same sentence its original paragraph ended. It talks about Isabel II's marriage with prince Filipe, their children, her father's death and her coronation.

Words that contributed with the header are highlighted in bold. In the first label, "Isabel" and "Elizabeth" were counted for "Isabel", with 8 repetitions. "Independentes", "independência", "soberano" and "sobreana" were counted for "Independente´' with 5 repetitions. "Estados" were counted towards "estado" with 4 repetitions. The word "rainha' occurred 3 times. "Reinos" counted towards "Reino", repeating twice.

For the second topic, only "ano" was repeated 2 times, considering the words "ano" and "anos". "Casou" and "príncipe" occurs only one time, counting for the labels "casar" and "príncipe", respectively. It is worth noting that the verb *"casar"* appears in its infinitive form due to the lemmatization process, while in the text segment it appears in its past tense *"casou"*.

3.3 Algorithm results

Table 3.2 shows the processing time in seconds for segmenting each text used in the user study. It is possible to see a connection between time and number of sentences. Texts 1 and 2 have similar values for time and number of sentences, and as Text 3 has almost double the sentences, its time triplicates. Comparing Text 3 with Text 4 also yields the same result, both time and sentences are doubled. This was expected, since the most expensive part of the algorithm is detecting topic boundaries between sentences in the rank matrix, which increases size depending on the number of sentences.

Table 3.2 – Time taken to segment each text

Text	Time(s)	Sentences	Words
Text 1	1,5	26	720
Text 2	1,26	28	748
Text 3	4,4	49	955
Text 4	10,47	82	1131

Source: from the author (2021)

In the case of the labelling algorithm (Table 3.3), it is the number of words that increases the processing time. It is possible to observe even though Text 3 has less topics than Texts 1 and 2, the time increases according to the number of words, and the same happens with Text 4. This was also expected, since the algorithm has to rank every word for each topic and create synonym sets to group words.

Table 3.3 – Time taken to label each text

Text	Time(s)	Topics	Words
Text 1	1,16	6	720
Text 2	1,13	6	748
Text 3	2,5	5	955
Text 4	3,21	7	1131

Source: from the author (2021)

Considering the aim of this tool is to be used in texts even longer than the ones in the user study, there is still much improvement needed to decrease processing time. For a text with just 1131 words, waiting 14 seconds to have automatically generated headers could discourage the user from using this tool. Conducting a user test with the algorithm implemented as a plugin would be necessary to confirm the impact of waiting time in the usability.

The segmentation algorithm was evaluated using WindowDiff (PEVZNER; HEARST, 2002), a measure for assessing text segmentation algorithms. This metric is used instead of precision or recall because it allows small differences in segmentation, *i.e.* the score suffers less impact when the topic boundary is close to where it should be. In our problem, segmentation does not need to be exact.

WindowDiff works as follows: the whole text is evaluated through a moving window that compares a sentence interval between the reference segmentation and the hypothetical–the one being evaluated. For each interval under evaluation, the amount of topical boundaries in the reference segmentation that are in that interval is compared with the amount in the automatic segmentation at the same interval. Intervals are scored as correct if they have the same number of boundaries between the beginning and end of the range. A corpus with texts separated by topic segments in Portuguese was necessary for the evaluation, the one chosen was CSTNews, a corpus with manually segmented journalistic texts. The result was an average **WindowDiff of 34.2%** across all texts in the corpus.

Using a multilingual BERT model allowed us to compare our results with previous research in English. We tested our algorithm with Choi dataset[6](CHOI, 2000) and measured error rate (P_k) (BEEFERMAN; BERGER; LAFFERTY, 1999). P_k is the probability that two randomly drawn sentences which are k sentences apart are classified incorrectly, lower values are better.

Table 3.4 compares the results we achieved with our algorithm and previous research. In the table we can see C99 and Misra et al. (2011) still outperforms our algorithm in most scenarios, except for the scenario with longest documents (9-11). Still, we were able detect boundaries with less errors than Reynar (REYNAR, 1998) and TextTiling (HEARST, 1997).

Table 3.4 – Performance comparison with previous algorithms

	3-5	6-8	9-11	3-11
Our Algorithm	27%	12%	8%	15%
C99	18%	10%	10%	13%
Reynar	21%	18%	16%	22%
Misra	16%	11%	11%	12%
TextTiling	44%	43%	48%	46%

Source: from the author (2021)

Although we based our algorithm on Naili's adaptation of C99 to use Word2Vec. It was not possible to compare results, since the corpus used by them was different and not made public.

The output of the automatic header generation algorithm is a text file with generated headers followed by their respective text segment. Each word from the header is separated by a space and is located right above its topic segment. Topics are separated by an empty line.

[6] Available in <http://web.archive.org/web/20010422042459/http://www.cs.man.ac.uk/~choif/software/C99-1.2-release.tgz>

4 USER EVALUATION OF AUTOMATICALLY GENERATED HEADERS TO NAVIGATE ON WEB PAGES

This chapter presents the methodology and results of the user study to evaluate a prototype of the automatic header generation algorithm. We gathered user feedback and measured the cognitive load and time to answer questions based on both texts without headers and with headers automatically generated by the proposed algorithm.

This study aimed to evaluate the effect of the generated headers in providing helpful information to the user to help them find the information they are seeking. The research was carried out with visually impaired users of screen readers to investigate aspects related to the understanding of automatically generated headers and the interaction of users with the proposed tool.

The project was analysed and approved by the research ethics committee on human beings from the Federal University of Lavras with protocol CAAE 45372321.3.0000.5148.

4.1 Methods for User Evaluation

This section describes the methods used for the user evaluation with the study design, participants recruitment and tasks. We describe the testing process to recruit participants and to make a form available online, simulating the final result of the algorithm.

4.1.1 Study Design

The study was conducted remotely due to the COVID-19 pandemic. The study involved evaluations with different web pages with four texts in two scenarios, in a mixed-design:

a) Without headers/ With headers: Texts 1 and 2 are presented in their original form, without headers, while texts 3 and 4 are presented with automatically generated headers by our algorithm;
b) With headers/ Without headers: Texts 1 and 2 are presented with automatically generated headers by our algorithm, while texts 3 and 4 are in their original form, without headers.

Participants were allocated for each group in turns when they volunteered, *i.e.* if a person was assigned to scenario a), the following volunteer would be assigned to scenario b). For each text, participants answered three questions before advancing to the subsequent text. If participants could not find an answer, they could skip it. After each text, participants rated the cognitive load used to complete the task, based on

Overall Workload (HILL et al., 1992). The tests were made available via Google Form, which also included terms of consent and demographic questions. Each session was recorded via Google Meet. The recordings were used to analyze users' interaction with long texts in both scenarios and identify possible problems.

4.1.2 Participants

All participants were at least 18 years old and had completed at least high school. Due to the limitations imposed by the COVID-19 pandemic, all tests were conducted remotely. All participants were required to have basic experience with screen readers and computers to complete the tasks since the researchers would not be able to provide in-loco assistance with technical difficulties. Table 4.1 shows demographic data of all participants.

Table 4.1 – Participants' Demographic data

Code	**Age**	**Education**	**Gender**	**Disability**	**Experience**	**Years using screen readers**
P1	26	Incomplete higher education	F	Total blindness	Intermediate	15
P2	33	Higher education	M	Total blindness	Intermediate	10
P3	51	Graduate degree	M	Residual light vision	Intermediate	20
P4	43	Graduate degree	M	Movement perception.	Intermediate	13
P5	28	Graduate degree	M	Total blindness	Advanced	12
P6	22	Higher education	M	Residual light vision	Advanced	16
P7	68	Higher education	M	Total blindness	Advanced	18
P8	28	Higher Education	F	Total blindness	Intermediate	20

Source: from the author (2021)

Participants were recruited by contacting people who participated in previous studies in the ALCANCE laboratory (Research Lab on Accessibility, Usability and Computational Linguistics) and posting an invitation for the study in Facebook groups aimed at visually impaired users.

4.1.3 Tasks

The study consisted of two scenarios, each with two tasks. For each task, participants had to answer a set of three questions based on the information provided by a long text.

To mitigate bias caused by unfamiliarity with the subject–participants could take a long time to understand something they are unfamiliar with – the texts were about different topics, and each task had two texts, totaling four texts used in the study. The texts are journalistic, extracted from the web, and expose a topic or explain a subject. To fit four texts in the user study without causing too much fatigue on volunteers, we limited texts to lengths between 720–1131 words (Appendix B). Information about the texts is presented as follows.

a) **Text 1**: *Novo coronavírus infecta e se replica em células das glândulas salivares – New coronavirus infects and replicates in salivary gland cells*
   - Text about a study to investigate if salivary glands are receptors of COVID-19.
   - Word count: 720.
   - Generated topics: 6
b) **Text 2**: *Covid: Risco de infecção por coronavírus varia 'muito' de acordo com máscara usada, diz estudo – Covid: Risk of coronavirus infection varies "a lot" by mask used, study says*
   - Text about a study conducted in United Kingdom to analyse the impact of using PFF3 face-masks in a hospital.
   - Word count: 748.
   - Generated topics: 6.
c) **Text 3**: *Por que a vida dos seus amigos parece ser melhor do que a sua? – Why do your friends' lives seem better than yours?*
   - Text about the *Paradox of Friendship*, which proposes that, on average, most people have fewer friends than their friends have.
   - Word count: 955.
   - Generated topics: 5.
d) **Text 4**: *Largada para o 2º Enem da pandemia – Start for the 2nd Enem of the pandemic*
   - Text about the application of the highschool national exam for the second time during the pandemic.
   - Word count: 1,131.

– Generated topics: 7.

There were no constraints on how participants should interact with the texts and questions. They could use any strategy and screen reader shortcuts they wanted, as well as the voice speed they preferred.

4.1.4 Test Procedure

Invitations were posted on Facebook groups related to visual impairment, and participants of previous studies in the ALCANCE lab were contacted. People who showed interest were contacted by e-mail or Whatsapp according to their preference, with further information related to the user study to schedule an online meeting to conduct the user test.

Each participant was assigned to one of the scenarios explained in the previous section when they agreed to participate. A link to the online meeting and the Google form for their respective scenario were also sent. The free and informed consent form (TCLE) was included in the form as the first page. The participant had to read and accept it before the study started. After accepting the TCLE, participants filled a demographic data form on the next page.

The tasks were composed of three questions and a text that could either have no headers or the automatically generated headers according to the scenario the participant was assigned to. They were allowed to interact with the text the way they preferred, with whichever voice speed, shortcuts or other techniques they were accustomed to. Participants were free to skip questions if unable to answer them, but everyone answered all questions.

After each task, participants evaluated the level of cognitive load they felt was necessary to complete the task. We used a scale from 1 (very low cognitive load) to 10 (very high cognitive load), based on Overall Workload (HILL et al., 1992). We favoured this scale because we wanted a simple way to measure mental effort for each task. We were aware of other alternatives to test cognitive load, such as the NASA-TLX (SWELLER; AYRES; KALYUGA, 2011) which separates the total cognitive load into six sub-scales, but we chose a more straightforward option to avoid tiredness that would impact the results of the following tasks. Hill et al. (1992) shows Overall Workload is both easier and faster to complete than the NASA-TLX. Gog et al. (2012) also showed evidence that repeatedly measuring mental effort after each task in the series is preferable to rating only once or retrospectively.

At the end of the test, questions were posed in a guided conversation to acquire feedback about the proposed algorithm. We preferred the conversation after the test because a method like *think-aloud*, where the user gives feedback while testing, would impact both the time and cognitive load measurement. The questions were separated into the ones about the headers and others about the topic segmentation as follows:

a) Do you think the tasks were easier with or without headers?

b) Did the headers help in finding information?

c) Could you infer from the headers what was inside your topics?

d) Do you have any other feedback related to the headers?

e) Did the topic segmentation help you to navigate through the texts?

f) What did you think about the size of the topics?

g) Did you have any difficulty?

h) Do you have any suggestion about what could be improved?

i) Would you use a tool like this if it was available for your screen reader?

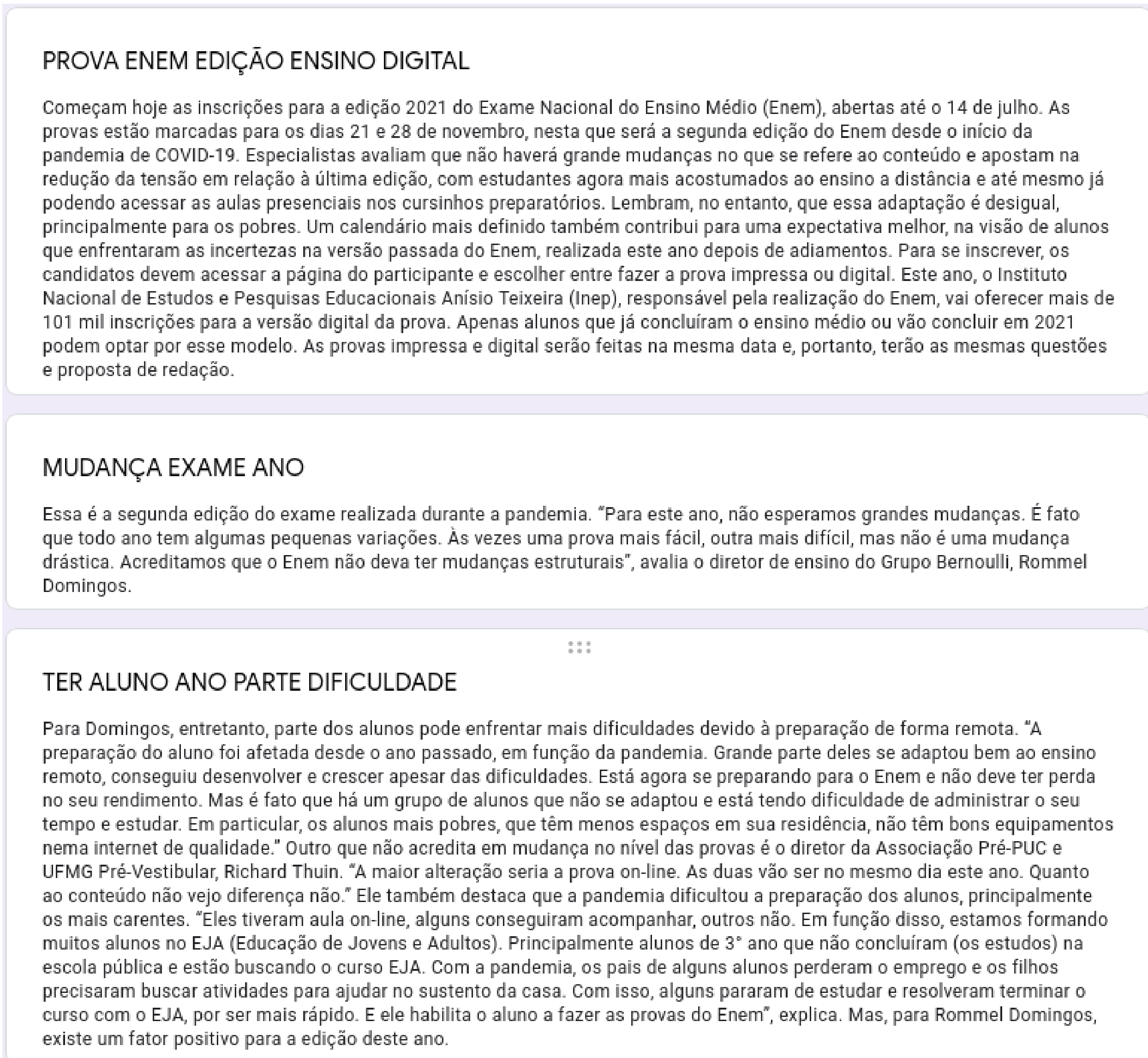
4.1.5 Prototype Evaluated

The prototype used in the user study comprises web pages built with the result of two algorithms, one to detect topic boundaries and divide the text accordingly and another to automatically label these topic segments by generating labels based on keywords from their respective text segment. The implementation of these algorithms is described in more detail in Chapter 3.

Four informative texts were selected to test our proposed algorithm, which divided and labelled topics automatically and saved them to a text file. Since the test was conducted remotely, it was needed to find a suitable way of making it available online while still allowing us to separate paragraphs with header and still be readable by a screen reader the same way a regular page would be. The platform chosen to create our forms was Google Form, since it is a well-known platform and allowed saving the results, sharing, and typical screen reader shortcuts worked in it. Figure 4.1 shows an example of a snippet of a text with automatically generated headers.

To simulate the proposed text result of our proposed algorithm inside Google Form, each topic was inserted in a separate text box to allow the use of both header and paragraph since Google Form does not allow for HTML editing. Each header is a *title* for a text box and the segment content is its *description*. We tested this in NVDA to ensure common shortcuts would work.

Figure 4.1 – Example of a text with generated headers



PROVA ENEM EDIÇÃO ENSINO DIGITAL

Começam hoje as inscrições para a edição 2021 do Exame Nacional do Ensino Médio (Enem), abertas até o 14 de julho. As provas estão marcadas para os dias 21 e 28 de novembro, nesta que será a segunda edição do Enem desde o início da pandemia de COVID-19. Especialistas avaliam que não haverá grande mudanças no que se refere ao conteúdo e apostam na redução da tensão em relação à última edição, com estudantes agora mais acostumados ao ensino a distância e até mesmo já podendo acessar as aulas presenciais nos cursinhos preparatórios. Lembram, no entanto, que essa adaptação é desigual, principalmente para os pobres. Um calendário mais definido também contribui para uma expectativa melhor, na visão de alunos que enfrentaram as incertezas na versão passada do Enem, realizada este ano depois de adiamentos. Para se inscrever, os candidatos devem acessar a página do participante e escolher entre fazer a prova impressa ou digital. Este ano, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), responsável pela realização do Enem, vai oferecer mais de 101 mil inscrições para a versão digital da prova. Apenas alunos que já concluíram o ensino médio ou vão concluir em 2021 podem optar por esse modelo. As provas impressa e digital serão feitas na mesma data e, portanto, terão as mesmas questões e proposta de redação.

MUDANÇA EXAME ANO

Essa é a segunda edição do exame realizada durante a pandemia. "Para este ano, não esperamos grandes mudanças. É fato que todo ano tem algumas pequenas variações. Às vezes uma prova mais fácil, outra mais difícil, mas não é uma mudança drástica. Acreditamos que o Enem não deva ter mudanças estruturais", avalia o diretor de ensino do Grupo Bernoulli, Rommel Domingos.

TER ALUNO ANO PARTE DIFICULDADE

Para Domingos, entretanto, parte dos alunos pode enfrentar mais dificuldades devido à preparação de forma remota. "A preparação do aluno foi afetada desde o ano passado, em função da pandemia. Grande parte deles se adaptou bem ao ensino remoto, conseguiu desenvolver e crescer apesar das dificuldades. Está agora se preparando para o Enem e não deve ter perda no seu rendimento. Mas é fato que há um grupo de alunos que não se adaptou e está tendo dificuldade de administrar o seu tempo e estudar. Em particular, os alunos mais pobres, que têm menos espaços em sua residência, não têm bons equipamentos nema internet de qualidade." Outro que não acredita em mudança no nível das provas é o diretor da Associação Pré-PUC e UFMG Pré-Vestibular, Richard Thuin. "A maior alteração seria a prova on-line. As duas vão ser no mesmo dia este ano. Quanto ao conteúdo não vejo diferença não." Ele também destaca que a pandemia dificultou a preparação dos alunos, principalmente os mais carentes. "Eles tiveram aula on-line, alguns conseguiram acompanhar, outros não. Em função disso, estamos formando muitos alunos no EJA (Educação de Jovens e Adultos). Principalmente alunos de 3° ano que não concluíram (os estudos) na escola pública e estão buscando o curso EJA. Com a pandemia, os pais de alguns alunos perderam o emprego e os filhos precisaram buscar atividades para ajudar no sustento da casa. Com isso, alguns pararam de estudar e resolveram terminar o curso com o EJA, por ser mais rápido. E ele habilita o aluno a fazer as provas do Enem", explica. Mas, para Rommel Domingos, existe um fator positivo para a edição deste ano.

Source: from the author (2021)

4.1.6 Data Analysis

We conducted a descriptive statistical analysis on cognitive load and time taken for task completion. The time was not considered when the study was disturbed by an external problem, such as the screen reader losing focus on the web page. We applied a within-subject paired-samples non- parametric Signed-Rank Wilcoxon test to assess whether there was a significant difference between texts with automatically generated headers and without headers for each participant.

We also conducted a qualitative analysis on the feedback collected during post-test guided interview, identifying positive and negative aspects of the proposed tool and situations that caused positive or negative reactions.

4.2 Results

This section presents the results obtained in the user study. Descriptive data for the time taken and cognitive load was measured to compare scenarios with and without automatically generated headers. Qualitative data was also obtained via a guided conversation with participants after finishing all tasks. We present their statements about the proposed tool, separated by topics.

4.2.1 Cognitive load

Cognitive load was measured to investigate if separating a long text into smaller topics labelled with automatically generated headers would reduce the mental effort for users. Based on the Overall Workload (HILL et al., 1992) measurement, a 10-point scale was used to measure cognitive load from 1–very low cognitive load to 10–very high cognitive load. This scale enabled measuring cognitive load with lower intervals, right after each task, without increasing the total testing time excessively.

Table 4.2 shows the cognitive load scores for each participant to complete tasks on each text. Texts with automatically generated headers are indicated with (h).

Table 4.2 – Cognitive load scores

	Scenario 1				Scenario 2			
	P1	P3	P5	P7	P2	P4	P6	P8
Text 1	4 (h)	2 (h)*	7 (h)	1 (h)	5	3	8	8**
Text 2	3 (h)	2 (h)	7 (h)	5 (h)	6	4	4	9*
Text 3	6	3	9	8	4 (h)	5 (h)*	6 (h)	6 (h)
Text 4	8	3	5	6	5 (h)	3 (h)	3 (h)	5 (h)
With headers	3,5	4,5	2	4	7	4,5	3	5,5
Without headers	7	5,5	3	3,5	7	6	7	8,5

*Tasks which had time deducted due to problems with the screen reader losing focus on the webpage.
**Task interrupted due the participant's notebook battery running out. The testing continued the next day.

Source: from the author (2021)

Since each scenario (with or without headers) had two texts, the average between each text for each scenario was calculated to test whether using automatically generated headers helped users navigate and find information inside texts.

A Signed-Rank Wilcoxon Test found no significant difference between the cognitive load of participants performing the task with or without headers (W = 16, N=8, p-value = 0.1). However, further studies with a larger number of participants would be necessary to analyze this hypothesis further.

4.2.2 Time on Task

The time for each task was measured from the moment the screen reader started reading the title until the participant clicked to enter the next page. Time taken with problems was deducted, *i.e.* screen reader losing focus from the web-page or the participant having to look for a charger. Time taken in each task was expected to vary between users since they were allowed to use their preferred voice speed. Table 4.3 shows the time taken in seconds for each participant to complete tasks on each text. Texts with automatically generated headers are indicated with (h).

Table 4.3 – Time taken on each task

	Scenario 1				Scenario 2			
	P1	P3	P5	P7	P2	P4	P6	P8
Text 1	539 (h)	531 (h)*	910 (h)	681 (h)	437	851	744	1570**
Text 2	402 (h)	451 (h)	572 (h)	878 (h)	454	473	574	1550*
Text 3	494	628	959	1418	640 (h)	432 (h)*	643 (h)	1069 (h)
Text 4	632	507	424	872	708 (h)	504 (h)	640 (h)	1166 (h)
With headers	471	473	741	780	674	468	642	1118
Without headers	563	568	692	1145	446	662	659	1560

*Tasks which had time deducted due to problems with the screen reader losing focus on the webpage.
**Task interrupted due the participant's notebook battery running out. The testing continued the next day.

Source: from the author (2021)

A Signed-Rank Wilcoxon Test found no significant difference between the time on task of participants performing the task with or without headers (W = 29, N=8, p-value = 0.79). This difference was not considered statistically significant despite indicating a possible reduction in time when comparing the mean times.

4.2.3 Participants' Perception on the Use of Automatically-Generated headers for Navigation

This section reports user's feedback on the use of the proposed automatically header generating algorithm for screen readers.

4.2.3.1 Usage of the headers

Four out of eight participants reported the headers facilitated the completion of tasks. This fact can be attributed to the segmentation of the text by topics and how the labels helped to know where participants were or infer the text's subject. The desired experience was achieved for some users, as Participant 2 reported "It was fantastic when I got used to it because I felt like running my eyes over the text.", almost like skimming the text.

For other users, the headers were helpful during a second read, helping locate the desired text segment. Participant 5 said *"It helps more for a second read when I've read it already, rather than in the first read"*. Participant 7 agreed the headers were helpful and said this tool would be even more useful for longer texts:

> *I certainly preferred [the texts] with markers [header]. And these texts weren't even that long. If they were longer it would be harder to remember where information was without the headers.*

In most cases, participants who did not use the generated headers explained they are used with line by line or paragraph reading to read the texts thoroughly and then answer the questions. As Participant 8 reported: *"In my case, it wasn't useful because I read it line by line."* However, even in this group, some participants considered the tool useful. As participant 4 said:

> *I didn't use the headers. I read paragraph by paragraph but if needed I could go back in the text at a certain point, headers would help.*

Participant 6 also said they did not use headers but argued:

> *It facilitated locating myself with the screen reader. In the texts with no headers I had to read line by line. In the ones with headers I just progressed reading the text and could go back to read afterwards.*

4.2.3.2 Infer topic content from header

Six out of eight participants reported being able to infer topic content based on the header. Participants 5 and 6 said they could not know the subject of a topic when reading the automatically generated header. Participant 5 reported "*the headers are a grouping of words that don't hold meaning.*" and that *"the words seemed disconnected"*. Participant 6 said the headers hinted at what the segment was about but was not enough to know for sure but also added that the text became easier to read:

> *It is noticeable that it helps finding information but I didn't use much. The words used in the headers can give us a hint, still, there is some doubt left. In the text without headers the impression is that it will never end. [...] Having the headers, with these highlighted words makes it less exhausting.*

Participant 6 also added headers could still improve in the choice of words and their presentation:

> *I think the algorithm can still improve on the choice of words. The words in the headers were from the text segment but the way they were placed, thrown in, could improve.*

The participants with positive responses said the labels were enough to guess somewhat the subject of its segment. Participant 2 said *"it helped to find what I was searching."* Participant 8 stated it was similar to using a strategy of searching for keywords: *"I think the headers were good according to what the text had to offer. Because many searches we do, we use keywords."*

4.2.3.3 Navigation

All participants said the topic separation helped to navigate in the text. Participant 1 answered positively, saying *"It was like navigating section by section"*, reminding the topic separations seen in Frequently Asked Questions of websites.

Participants 4, 5, 7 also reported that the headers helped them to remember where something is and find it again. Participant 5 said the tool was a positive surprise because at first, they were not convinced about it, but after using it helped to locate themselves when navigating the text for question answering:

> *I started thinking it [the headers] was unnecessary, [...] but by simply having them it created a checkpoint for me. [...] It helps me to fragment the text in my mental reading. I can get used to memorizing where something is by its header.*

Participant 6 said the headers helped navigating for another reason, it enabled him to use the header listing shortcut to navigate the text:

> *It helped a lot because I could use shift+H to navigate through headers, it was much faster. It gives me agility.*

4.2.3.4 Desire to use the tool

All participants said they would use a tool like this, if available. Participant 6 stated it would be a useful tool since *"There are many sites that don't even have headers."*. Participant 5 was eager to use an algorithm like this for his work: *"Yes, mainly for my reading activity, proofreading."* Participant 2 also commented on a problem they often face when reading long texts with no headers:

> *When there is no marking, navigating takes more time because we have no anchor or reference to use with the screen reader.*

Two participants reported the segmentation did not seem to have been done by the machine that they felt like topics were breaking where they should, and it looked like the algorithm was breaking texts at the end of paragraphs. However, it is worth noting that when the segmenting algorithm process the input text, it has no information about paragraphs. It simply processes a list of sentences. Participant 2 said:

> *In most cases I didn't notice it was made by machine. In some points I thought 'this is a little weird' because the subject continued in the next paragraph.*

Participant 5 also reported something similar:

> *In some cases I felt like the topics were too much divided by paragraphs, and not by subject. The algorithm would do 'this paragraph is this block, this other paragraph is that block'. [...] I felt like the topic breaks were actually the paragraphs, just like someone had intentionally ended one and started the next. If this was the algorithm, it's very good.*

Because of these statements, we checked the number of paragraphs from the original text that matched the topic boundaries detected by our algorithm. Text 1 had five boundaries; 4 of them were at the end of a paragraph. On text 2, all boundaries matched the ends of paragraphs. For text 3, there were 3 boundaries out of 4 that matched paragraph breaks. On text 4, just two boundaries out of 6 matched the end of a paragraph. This might indicate that, even in cases where the boundaries were not in the same place as the paragraph ending, it still made enough sense for the participant to think it was a natural break.

5 DISCUSSION

The objective of this project was to implement and evaluate a tool to automatically generate headers in long texts in Portuguese, which could be adapted as a plugin for screen readers to aid visually impaired people in information-seeking tasks. An algorithm for automatically segmenting tasks based on topic shift and automatically generating headers based on keywords of these topic segments was proposed to achieve this goal. Both algorithms can work together to allow screen reader users to navigate inside a document based on the topic they are seeking. The proposed algorithm was evaluated through a user study with 8 participants, all screen reader users.

5.1 Analysis of the algorithms

In this Section we discuss how our algorithms performed, comparing with previous studies and the implications of our results. In this project we implemented a topic segmentation algorithm adapting C99 to use BERT, and a topic labelling algorithm based on keywords. In conjunction, these algorithms enabled us to automatically generate headers

We experimented with BERT in the topic segmenting algorithm to observe how this state-of-the-art model would improve an already established algorithm, C99. This decision was inspired by Naili, Chaibi and Ben Ghezala (2017), which used Word2Vec and LSA to calculate sentence similarities. We first experimented using different layers of a BERT model to calculate cosine similarities, but the performance was slow. Speed was a vital aspect of this algorithm since screen reader users would have to wait for the process to finish to have the headers available. We then experimented with SBERT, which uses the GPU in the similarity calculation process, speeding up the process.

Our results indicate that the proposed algorithm did not improve the original algorithm, C99. It had a higher error rate on detecting topic boundaries in most cases, which is worse. The only scenario where our algorithm yielded a better error rate than C99 was with the longest documents subset (9-11). This is a positive indicator since our project aimed to segment long texts for visually impaired users. Still, the time it takes to finish the segmentation process would heavily affect how users perceive this tool.

Further research can be done to assess if the error rate would improve for even longer documents, but time is a considerable limitation for our application. It also might be fruitful to try fine-tuning a BERT model specifically for the topic segmenting task in Portuguese and compare if it improves results. Nevertheless, in order to do that, a corpus with long segmented documents in Portuguese is necessary.

For the labelling algorithm, we first experimented with Bhatia, Lau and Baldwin (2016), which used both Word2Vec and Doc2Vec trained in a Wikipedia dump to compare the top-10 words for each topic and relate it to an article title from Wikipedia. Our results were highly unsatisfactory with headers that had nothing in common with the text subject. We suppose this is because, instead of labelling a document, we tried using this algorithm to label text segments, which were much smaller and lacked enough information. Another problem we had with this algorithm for this task is that even when titles were related to the topic, they were still inadequate, limiting the meaning too much or making it too broad. This fact indicated that even if we improved the algorithm, it would still be inadequate for our task. Hence, we needed to develop a new one. We decided on using a statistical approach to label topic segments.

Even though machine learning algorithms can achieve impressive results, it is still necessary to have a corpus big enough to be correctly trained. Furthermore, this corpus can be task-dependent. In the case of our problem, a corpus with segmented and labelled texts for Portuguese would be needed. Even if we found it, when results are unsatisfactory, trying to pinpoint what influenced the result proved to be hard when we tried the Wikipedia-based label algorithm. The time necessary to train a new model also needs to be taken into account. For these reasons, we decided to focus the machine learning process on the segmenting algorithm and use a statistical algorithm for the labelling step.

5.2 User Study Implications

Although the number of participants in the user study was insufficient to determine whether the proposed tool positively impacted participants, we can see an indication that cognitive load and time measurements did decrease in texts with headers. Not only that, but users also reported they felt it was easier to complete tasks with automatically generated headers.

One interesting finding is that text segmentation seemed to be more helpful for the users than the headers themselves. Half of the participants reported not using the headers when completing the tasks, but even they agreed that having headers separating the text by topic helped navigate, locate themselves, and refind information in a second read. According to participants, this fact made the tasks with headers less exhaustive, indicating how impactful navigation tools inside texts can be helpful. Participants considering segmented texts easier to navigate is in line with what Tombaugh, Lickorish and Wright (1987) observed, that smaller fragments of texts–in our case topic segments–help locate portions of previously read texts by providing spatial cues, which aids in recollecting information.

Participant 5 detailed why navigation tools are important, reading long texts in their entirety would take too much time, and be very exhaustive when using a screen reader. He then described strategies applied when reviewing texts. One of such strategies is reading the start of paragraphs when looking for information because it often summarizes its subject, though this is not always the case. Another strategy for reading texts 70 pages long is sampling segments: reading the first, second paragraph, then the sixth, tenth, jumping through intervals to get a general sense of meaning. He also expressed discontent for screen readers not having the option to navigate by sentence. He either reads from the start of a paragraph or at the start of each line. However, he cannot choose a sentence in the middle of a paragraph.

Participant 7 also reported a different strategy when reading long texts. He called it a *dynamic reading*: speeding up the screen reader to read the text fast without paying much attention and picking up keywords or concepts. Then, he searches for the keywords to read, paying more attention. This resembles the *Discovery* strategy pointed out by Power et al. (2013), where the user gains an overview of a web-page. But in this case, he applied it specifically for the text to locate topics of interest.

As observed by Borodin et al. (2010), common strategies to skip content participants included using header navigation shortcuts and keyword search, both were used for a second read. But unlike what they observed, participants used sequential navigation as a first resort, skipping repeated content (such as the title of the research in every page). This corroborates to the difference we observed when comparing research about screen reader usage worldwide (WEBAIM, 2021) and in Brazil (EVERIS, 2020). Brazilians tend to read through the page much more than the rest of the world, where heading navigation is the norm.

In three instances we observed what Power et al. (2013) called *Reset* strategy. When a user abandons what they were doing and starts again at a safe point in order to work around an error. This happened twice when the screen reader lost focus on the web-page and got stuck in the navigation bar area. The user was unable to figure out how to return to the page so he restarted the browser. In another instance, Windows Update popped-up, confusing the user, who decided to close every window and start again.

Although some users reported disliking the way words were presented, or even the choice of words to be used as a label, most users stated being able to infer the subject of topics by reading the label. We acknowledge there is still room for improvement in the labels, such as using syntax trees or language models when choosing words, but it was interesting to see this positive result. Participant 8 indicated why the labels were intuitive. Despite being just a set of words, blind users often use keywords to find what they are looking for on a web-page. Participant 2 even reported feeling like he was “running his eyes through the text”, our initial inspiration for this project, recreating text skimming for visually impaired people.

5.3 The Viability of Implementing the Proposal as a Screen Reader Plugin

The aim of the proposed algorithm is to be used with screen readers. Thus, we aimed to make it simple to adapt into a plugin for mainstream screen reader software, such as NVDA. The input for the algorithm is a list of sentences, and the developer can use whatever methods they prefer to extract the text from a web-page to serve as input for the algorithm. In our case, this was done by using a Python library called jusText to extract the main content of HTML pages. The output by default is a text document with headers and their related topic segment, but this behaviour can be changed since sentences, topic segmentation, and related headers are stored in separate variables.

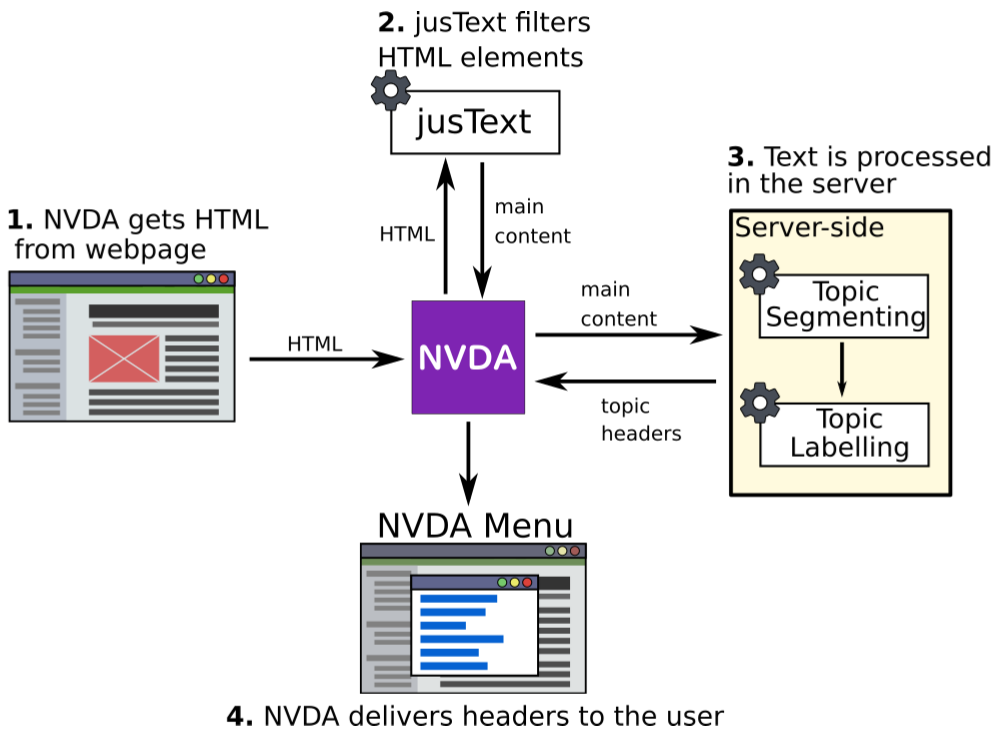
One limitation of the proposed algorithm when implementing it in a plugin is that it must be processed in an external server. Since a BERT model is needed for the segmentation process, the user would need to download it. The model used in this project is 482MB, quite big for a plugin. Not only that, but the algorithm also consumed up to 2GB of RAM during the processing of the matrix. This would have a heavy impact on the user's computer performance if implemented client-side. There are lighter alternatives to BERT, such as ALBERT (LAN et al., 2019), but at the time of the algorithm development we were unable to find a Portuguese trained model. We acknowledge at the time of writing that an ALBERT model trained in Portuguese is now available (FEIJO; MOREIRA, 2020).

Figure 5.1 illustrates an example framework for implementing this algorithm into a screen reader. In this example, the algorithm is being used to generate headers in a web-page automatically. Screen readers have a method for accessing the HTML content (1) which can be filtered by jusText (2) to create an input text containing only the primary information on the page, *e.g.* the body of an article. This document is processed server-side (3) with the topic segmentation algorithm to create a list of topic segments that are then separately labelled by the topic labelling algorithm, resulting in a list of labelled topic segments. The result can be integrated with the screen reader through a menu (4) containing the list of generated topics or by injecting the web page's headers.

5.4 Limitations

The main limitation of this work is that the user study was not conducted with the algorithm implemented as a plugin in a screen reader. The test was conducted using preprocessed text, which enabled us to evaluate how participants interacted with the automatically generated headers, analysing their impact on

Figure 5.1 – Framework for using this addon as a plugin



Source: From the author (2021)

information seeking tasks and gathering helpful feedback. Nonetheless, without using an actual plugin with the screen reader, we cannot evaluate user experience with the tool. Many factors could interfere when using an actual plugin, such as default shortcuts, the time to process the text and generate headers, and responsiveness. This fact is a limitation our work shares with previous research (AHMED et al., 2013; ALVES; CARDOSO; FREIRE, 2018).

The COVID pandemic caused another limitation we faced. User tests had to be conducted remotely, which forced us to recruit participants with experience in using computers and screen readers, the least experienced participant used screen readers for ten years at the time of the study. This fact influences results since users with basic computer knowledge or beginner screen-reader users might encounter other problems and give further feedback.

Even though participants were experienced screen reader users, the tests were not without problems. When doing tests, the main issue was the screen reader occasionally changing the focus from the web page to the address bar. This problem can generally happen when reading web pages, not being specific to our test. There was also one occurrence of Windows Update popping up and stealing the focus from the web-page, which confused the participant, disturbing the testing process for a minute.

The number of participants (8) was too small to cover all possible combinations between texts and scenarios (texts with and without automatically generated headers). However, scenarios were alternated between participants to diminish the influence in the results.

Due to time restrictions for an online test–the total time for each test ranged from 43 minutes up to 2 hours–we had to compromise and choose texts smaller than our initial target to avoid fatigue for participants. Future works in the laboratory, done in more than a day, could explore longer texts.

Another limitation caused by the online application of this study were external factors, such as noises, interruptions or conversations. These factors are difficult to control even in a laboratory, but with participants being in their homes, distractions could and did occur.

Our results showed that, although users saw the benefit of using the proposed tool, the labels for the headers still need to be improved. They need improvements both in terms of how they are presented, and in the choice of words. In the present state, there is no document-wide awareness, as each topic is labeled separately. This causes repetitive words through the document to appear in many headers, not adding value in terms of differentiating one segment to the other.

6 CONCLUSION AND FUTURE WORK

The objective of this study was to investigate whether automatic header generation can be used to improve screen reader users' usability in information seeking tasks in web content.

To achieve this goal, this study involved a systematic mapping process to synthesize previous work on improving usability with screen readers, the proposal and implementation of an algorithm to automatically generate headers for large documents and its evaluation via a user study with screen reader users.

We developed an automatic header generation algorithm using text segmentation and topic labelling algorithms. We experimented with state-of-the-art word embedding models to test if they could improve previous results. The results showed that for most cases, our implementation of C99 with BERT failed to improve other results in the literature. However, for longer documents, which was our focus, we achieved better results when compared with previous research.

A user study with 8 participants indicated that the segmentation process was good enough for users to benefit from using the tool. Participants had positive reactions when interacting with texts processed by our proposed algorithm, and they reported that the topic segments separation helped locate information and navigate the text.

Most participants also reported being able to infer the subject of a text segment by its label and that the headers helped when recollecting information in a second read. Other participants argued that the choice of words and the way they are present could be improved. Even so, all users were eager to have a similar tool available for their screen readers to have more navigation options.

Still, there is room for improvement. Future works could explore different segmenting and labelling techniques. Our results were too slow to be implemented client-side, and word embedding models proved to have limitations for this problem. A corpus of segmented texts in Portuguese would benefit from training models or fine-tuning BERT specifically for this task. In order to be applied as an actual plugin, the tool also needs to be faster since waiting times would impact usability.

The selection of words for the labels can also improve. We did not use word-sense disambiguation in this project, which caused synsets to be too broad at times. Our algorithm also does not weight a word's repetition through the whole text, and this causes labels to sometimes repeat words across the text. Using tf-idf could prevent this. Some users also commented that the labels seemed like a bunch of words thrown together. Applying concepts such as syntax trees can improve the choice of words to form a more meaningful label.

The COVID pandemic also limited our options for the user study. Future work could involve participants of different ranges of experience with screen readers or computers. It would also be beneficial to implement the algorithm as a plugin, allowing to collect feedback about user interaction with the plugin. Testing with longer documents is also important. We needed to choose texts shorter than initially planned due to the difficulties of running a user study remotely. In a in-person user study, researchers can have higher control of external impacts on the evaluation.

BIBLIOGRAPHY

AHMED, F. et al. Accessible skimming: Faster screen reading of web pages. In: **Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology**. New York, NY, USA: Association for Computing Machinery, 2012. (UIST '12), p. 367–378. ISBN 9781450315807. Available from Internet: <https://doi.org/10.1145/2380116.2380164>.

AHMED, F. et al. Non-visual skimming on touch-screen devices. In: **Proceedings of the 2013 International Conference on Intelligent User Interfaces**. New York, NY, USA: ACM, 2013. (IUI '13), p. 435–444. ISBN 978-1-4503-1965-2. Available from Internet: <http://doi.acm.org/10.1145/2449396.2449452>.

ALEMI, A. A.; GINSPARG, P. Text segmentation based on semantic word embeddings. **arXiv preprint arXiv:1503.05543**, 2015. Available from Internet: <http://arxiv-export-lb.library.cornell.edu/abs/1503.05543?context=cs>.

ALTSZYLER, E. et al. Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. **arXiv preprint arXiv:1610.01520**, 2016. Available from Internet: <https://arxiv.org/abs/1610.01520>.

ALVES, E. A.; CARDOSO, P. C. F.; FREIRE, A. P. Automatically generated summaries as in-page web navigation accelerators for blind users. In: **Proceedings of the 17th Brazilian Symposium on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2018. (IHC 2018). ISBN 9781450366014. Available from Internet: <https://doi.org/10.1145/3274192.3274202>.

BARONI, M.; DINU, G.; KRUSZEWSKI, G. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Baltimore, Maryland: Association for Computational Linguistics, 2014. p. 238–247. Available from Internet: <https://aclanthology.org/P14-1023>.

BARZILAY, R.; MCKEOWN, K. R.; ELHADAD, M. Information fusion in the context of multi-document summarization. In: **Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics**. Stroudsburg, PA, USA: Association for Computational Linguistics, 1999. (ACL '99), p. 550–557. ISBN 1-55860-609-3. Available from Internet: <https://doi.org/10.3115/1034678.1034760>.

BASAVE, A. E. C.; HE, Y.; XU, R. Automatic labelling of topic models learned from twitter by summarisation. In: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Baltimore, Maryland: Association for Computational Linguistics, 2014. p. 618–624. Available from Internet: <https://www.aclweb.org/anthology/P14-2101>.

BEEFERMAN, D.; BERGER, A.; LAFFERTY, J. Statistical models for text segmentation. **Machine learning**, Springer, v. 34, n. 1, p. 177–210, 1999. Available from Internet: <https://doi.org/10.1023/a:1007506220214>.

BHATIA, S.; LAU, J. H.; BALDWIN, T. Automatic labelling of topics with neural embeddings. In: **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**. Osaka, Japan: The COLING 2016 Organizing Committee, 2016. p. 953–963. Available from Internet: <https://www.aclweb.org/anthology/C16-1091>.

BOJANOWSKI, P. et al. Enriching word vectors with subword information. **Transactions of the Association for Computational Linguistics**, MIT Press, v. 5, p. 135–146, 2017. Available from Internet: <https://aclanthology.org/Q17-1010>.

BORODIN, Y. et al. More than meets the eye: A survey of screen-reader browsing strategies. In: **Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)**. New York, NY, USA: Association for Computing Machinery, 2010. (W4A '10). ISBN 9781450300452. Available from Internet: <https://doi.org/10.1145/1805986.1806005>.

BRASIL. Decreto legislativo n.^{o} 186, de 9^{o} de julho de 2008. Aprova o texto da Convenção sobre os Direitos das Pessoas com Deficiência e de seu Protocolo Facultativo, assinados em Nova Iorque, em 30 de março de 2007. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 9 de jul. 2008. Available from Internet: <http://www.planalto.gov.br/ccivil_03/CONGRESSO/DLG/DLG-186-2008.htm>. Cited 29 jan. 2020.

BRASIL. Lei n.^{o} 13.146, de 6^{o} de julho de 2015. Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência). **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 6 de jul. 2015. Available from Internet: <http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2015/Lei/L13146.htm>. Cited 29 jan. 2020.

BROOKE, J. et al. Sus-a quick and dirty usability scale. **Usability evaluation in industry**, London, v. 189, n. 194, p. 4–7, 1996.

BROWN, G. et al. Discourse analysis. **Journal of Linguistics**, Cambridge university press, p. 241–245, 1983.

BUCY, E. P.; NEWHAGEN, J. E. **Media access: Social and psychological dimensions of new technology use**. New Jersey, EUA: Psychology Press, 2004.

CARVER, R. P. Rauding theory predictions of amount comprehended under different purposes and speed reading conditions. **Reading Research Quarterly**, JSTOR, v. 19, n. 2, p. 205, 1984. Available from Internet: <https://doi.org/10.2307/747363>.

CHANDLER, P.; SWELLER, J. Cognitive load theory and the format of instruction. **Cognition and Instruction**, Routledge, v. 8, n. 4, p. 293–332, 1991. Available from Internet: <https://doi.org/10.1207/s1532690xci0804_2>.

CHERRY, E. C. Some experiments on the recognition of speech, with one and with two ears. **The Journal of the Acoustical Society of America**, v. 25, n. 5, p. 975–979, 1953. Available from Internet: <https://doi.org/10.1121/1.1907229>.

CHOI, F. Y. Y. Advances in domain independent linear text segmentation. In: **Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference**. USA: Association for Computational Linguistics, 2000. (NAACL 2000), p. 26–33. Available from Internet: <https://dl.acm.org/doi/10.5555/974305.974309>.

CHOI, F. Y. Y.; WIEMER-HASTINGS, P.; MOORE, J. Latent semantic analysis for text segmentation. In: **Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing**. [s.n.], 2001. Available from Internet: <https://www.aclweb.org/anthology/W01-0514>.

CONNEAU, A. et al. Supervised learning of universal sentence representations from natural language inference data. Association for Computational Linguistics, Copenhagen, Denmark, p. 670–680, set. 2017. Available from Internet: <https://aclanthology.org/D17-1070>.

DEERWESTER, S. et al. Indexing by latent semantic analysis. **Journal of the American Society for Information Science**, v. 41, n. 6, p. 391–407, 1990. Available from Internet: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9>.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. Association for Computational Linguistics, Minneapolis, Minnesota, p. 4171–4186, jun. 2019. Available from Internet: <https://aclanthology.org/N19-1423>.

DUGGAN, G. B.; PAYNE, S. J. Text skimming: The process and effectiveness of foraging through text under time pressure. **Journal of Experimental Psychology: Applied**, American Psychological Association (APA), v. 15, n. 3, p. 228–242, 2009. Available from Internet: <https://doi.org/10.1037/a0016995>.

DYSON, M.; HASELGROVE, M. The effects of reading speed and reading patterns on the understanding of text read from screen. **Journal of Research in Reading**, Wiley, v. 23, n. 2, p. 210–223, jun. 2000. Available from Internet: <https://doi.org/10.1111/1467-9817.00115>.

EISENSTEIN, J.; BARZILAY, R. Bayesian unsupervised topic segmentation. In: **Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing**. Honolulu, Hawaii: Association for Computational Linguistics, 2008. p. 334–343. Available from Internet: <https://aclanthology.org/D08-1035>.

EVERIS. **2ª pesquisa brasileira do uso de leitores de tela**. 2020. Available from Internet: <https://estudoinclusivo.com.br/pesquisa-ldt-2019-relatorio.pdf>.

FEIJO, D. d. V.; MOREIRA, V. P. Mono vs multilingual transformer-based models: a comparison across several language tasks. **arXiv preprint arXiv:2007.09757**, 2020. Available from Internet: <https://arxiv.org/abs/2007.09757>.

FERRI, D.; FAVALLI, S. Web accessibility for people with disabilities in the european union: Paving the road to social inclusion. **Societies**, Multidisciplinary Digital Publishing Institute, v. 8, n. 2, 2018. ISSN 2075-4698. Available from Internet: <https://www.mdpi.com/2075-4698/8/2/40>.

GADDE, P.; BOLCHINI, D. From screen reading to aural glancing: Towards instant access to key page sections. In: **Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility**. New York, NY, USA: Association for Computing Machinery, 2014. (ASSETS ’14), p. 67–74. ISBN 9781450327206. Available from Internet: <https://doi.org/10.1145/2661334.2661363>.

GHAHARI, R. R. et al. Back navigation shortcuts for screen reader users. In: **Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility**. New York, NY, USA: Association for Computing Machinery, 2012. (ASSETS ’12), p. 1–8. ISBN 9781450313216. Available from Internet: <https://doi.org/10.1145/2384916.2384918>.

GIRAUD, S.; THÉROUANNE, P.; STEINER, D. D. Web accessibility: Filtering redundant and irrelevant information improves website usability for blind users. **International Journal of Human-Computer Studies**, v. 111, p. 23 – 35, 2018. ISSN 1071-5819. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1071581917301490>.

GOG, T. van et al. Timing and frequency of mental effort measurement: Evidence in favour of repeated measures. **Applied Cognitive Psychology**, Wiley Online Library, v. 26, n. 6, p. 833–839, 2012. Available from Internet: <https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.2883>.

GUERREIRO, J.; GONÇALVES, D. Faster text-to-speeches: Enhancing blind people's information scanning with faster concurrent speech. In: **Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility**. New York, NY, USA: ACM, 2016. (ASSETS '15), p. 3–11. ISBN 978-1-4503-3400-6. Available from Internet: <http://doi.acm.org/10.1145/2700648.2809840>.

GUERREIRO, J. a.; GONÇALVES, D. Text-to-speeches: Evaluating the perception of concurrent speech by blind people. In: **Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility**. New York, NY, USA: Association for Computing Machinery, 2014. (ASSETS '14), p. 169–176. ISBN 9781450327206. Available from Internet: <https://doi.org/10.1145/2661334.2661367>.

GUERREIRO, J. a.; GONÇALVES, D. Scanning for digital content: How blind and sighted people perceive concurrent speech. Association for Computing Machinery, New York, NY, USA, v. 8, n. 1, 2016. ISSN 1936-7228. Available from Internet: <https://doi.org/10.1145/2822910>.

HART, S. G.; STAVELAND, L. E. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In: HANCOCK, P. A.; MESHKATI, N. (Ed.). **Human Mental Workload**. North-Holland, 1988, (Advances in Psychology, v. 52). p. 139 – 183. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0166411508623869>.

HEARST, M. A. Texttiling: Segmenting text into multi-paragraph subtopic passages. **Computational Linguistics**, MIT Press, Cambridge, MA, USA, v. 23, n. 1, p. 33–64, mar. 1997. ISSN 0891-2017. Available from Internet: <http://dl.acm.org/citation.cfm?id=972684.972687>.

HILL, S. G. et al. Comparison of four subjective workload rating scales. **Human Factors**, Human Factors & Ergonomics Society, Inc., USA, v. 34, n. 4, p. 429–439, ago. 1992. ISSN 0018-7208. Available from Internet: <https://doi.org/10.1177/001872089203400405>.

IBGE. **Censo Demográfico 2010: Características gerais da população, religião e pessoas com deficiência**. Rio de Janeiro: IBGE, 2011. Available from Internet: <https://biblioteca.ibge.gov.br/index.php/biblioteca_catalogo?view=detalhes&id=794>.

JAIN, M.; DIWAKAR, N.; SWAMINATHAN, M. Smartphone usage by expert blind users. In: **Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2021. ISBN 9781450380966. Available from Internet: <https://doi.org/10.1145/3411764.3445074>.

KLEIN, D.; MANNING, C. D. Accurate unlexicalized parsing. In: **Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1**. USA: Association for Computational Linguistics, 2003. (ACL '03), p. 423–430. Available from Internet: <https://doi.org/10.3115/1075096.1075150>.

KOU, W.; LI, F.; BALDWIN, T. Automatic labelling of topic models using word vectors and letter trigram vectors. In: **Information Retrieval Technology**. Cham: Springer International Publishing, 2015. p. 253–264. ISBN 978-3-319-28940-3.

KOZIMA, H. Text segmentation based on similarity between words. In: **Proceedings of the 31st Annual Meeting on Association for Computational Linguistics**. USA: Association for Computational Linguistics, 1993. (ACL '93), p. 286–288. Available from Internet: <https://doi.org/10.3115/981574.981616>.

LAN, Z. et al. ALBERT: A lite BERT for self-supervised learning of language representations. **CoRR**, abs/1909.11942, 2019. Available from Internet: <http://arxiv.org/abs/1909.11942>.

LANDAUER, T. K.; DUMAIS, S. T. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. **Psychological review**, American Psychological Association, v. 104, n. 2, p. 211, 1997.

LAU, J. H. et al. Automatic labelling of topic models. In: **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**. Portland, Oregon, USA: Association for Computational Linguistics, 2011. p. 1536–1545. Available from Internet: <https://www.aclweb.org/anthology/P11-1154>.

LAU, J. H. et al. Best topic word selection for topic labelling. In: **Proceedings of the 23rd International Conference on Computational Linguistics: Posters**. USA: Association for Computational Linguistics, 2010. (COLING '10), p. 605–613. Available from Internet: <https://dl.acm.org/doi/10.5555/1944566.1944635>.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: **Proceedings of the 31st International Conference on International Conference on Machine Learning**. Journal of Machine Learning Research, 2014. (ICML'14, v. 32), p. 1188—1196. Available from Internet: <https://dl.acm.org/doi/10.5555/3044805.3045025>.

LI, Q. et al. PAR: Partition-annotate-recommend paradigm for improved accessibility in digital libraries. In: IEEE. **2010 International Conference on Management of e-Commerce and e-Government**. 2010. p. 3–9. Available from Internet: <https://doi.org/10.1109/icmecg.2010.10>.

MAGATTI, D. et al. Automatic labeling of topics. In: **2009 Ninth International Conference on Intelligent Systems Design and Applications**. IEEE, 2009. Available from Internet: <https://doi.org/10.1109/isda.2009.165>.

MANISHINA, E. et al. Tag thunder: Towards non-visual web page skimming. In: **Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility**. New York, NY, USA: ACM, 2016. (ASSETS '16), p. 281–282. ISBN 978-1-4503-4124-0. Available from Internet: <http://doi.acm.org/10.1145/2982142.2982152>.

MAO, X.-L. et al. Automatic labeling hierarchical topics. In: **Proceedings of the 21st ACM International Conference on Information and Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2012. (CIKM '12), p. 2383–2386. ISBN 9781450311564. Available from Internet: <https://doi.org/10.1145/2396761.2398646>.

MASSON, M. E. Cognitive processes in skimming stories. **Journal of Experimental Psychology: Learning, Memory, and Cognition**, American Psychological Association (APA), v. 8, n. 5, p. 400–417, set. 1982. Available from Internet: <https://doi.org/10.1037/0278-7393.8.5.400>.

MAZIERO, E. G. et al. A base de dados lexical e a interface web do tep 2.0: Thesaurus eletrônico para o português do brasil. In: **Companion Proceedings of the XIV Brazilian Symposium on Multimedia and**

the Web. New York, NY, USA: Association for Computing Machinery, 2008. (WebMedia ’08), p. 390–392. ISBN 9788576691990. Available from Internet: <https://doi.org/10.1145/1809980.1810076>.

MEI, Q.; SHEN, X.; ZHAI, C. Automatic labeling of multinomial topic models. In: **Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2007. (KDD ’07), p. 490–499. ISBN 9781595936097. Available from Internet: <https://doi.org/10.1145/1281192.1281246>.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: **Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2**. Red Hook, NY, USA: Curran Associates Inc., 2013, (NIPS’13). p. 3111–3119. Available from Internet: <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>.

MILLER, G. A. **WordNet: An electronic lexical database**. Cambridge, MA ; London: MIT press, 1998. Available from Internet: <https://doi.org/10.7551/mitpress/7287.001.0001>.

MISRA, H. et al. Text segmentation: A topic modeling perspective. **Information Processing & Management**, v. 47, n. 4, p. 528–544, 2011. ISSN 0306-4573. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0306457310000981>.

MORENO, L. et al. Support resource based on standards for accessible e-government transactional services. **Computer Standards & Interfaces**, Elsevier Science Publishers B. V., NLD, v. 58, n. C, p. 146–157, maio 2018. ISSN 0920-5489. Available from Internet: <https://doi.org/10.1016/j.csi.2018.01.003>.

NAILI, M.; CHAIBI, A. H.; Ben Ghezala, H. H. Comparative study of word embedding methods in topic segmentation. **Procedia Computer Science**, Marseille, France, v. 112, p. 340–349, 2017. ISSN 1877-0509. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S1877050917313480>.

NAILI, M.; HABACHA, A. C.; GHEZALA, H. H. B. Parameters driving effectiveness of lsa on topic segmentation. In: **Computational Linguistics and Intelligent Text Processing**. Cham: Springer International Publishing, 2018. p. 560–572. ISBN 978-3-319-75477-2. Available from Internet: <https://link.springer.com/chapter/10.1007\%2F978-3-319-75477-2_40>.

NENKOVA, A.; VANDERWENDE, L. The impact of frequency on summarization. **Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005**, v. 101, 2005. Available from Internet: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.529.6099>.

PAGE, L. et al. **The PageRank Citation Ranking: Bringing Order to the Web.** California, USA, 1999. Available from Internet: <http://ilpubs.stanford.edu:8090/422/>.

PECINA, P. Lexical association measures and collocation extraction. **Language Resources and Evaluation**, Springer Science and Business Media LLC, v. 44, n. 1-2, p. 137–158, out. 2009. Available from Internet: <https://doi.org/10.1007/s10579-009-9101-4>.

PERNICE, K. Banner blindness revisited: users dodge ads on mobile and desktop. **Nielsen Norman Group**, 2018. Available from Internet: <https://www.nngroup.com/articles/banner-blindness-old-and-new-findings/>.

PETERSEN, K. et al. Systematic mapping studies in software engineering. In: **Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering**. Swindon, UK: BCS Learning & Development Ltd., 2008. (EASE'08), p. 68–77. Available from Internet: <http://dl.acm.org/citation.cfm?id=2227115.2227123>.

PEVZNER, L.; HEARST, M. A. A critique and improvement of an evaluation metric for text segmentation. **Computational Linguistics**, MIT Press, Cambridge, MA, USA, v. 28, n. 1, p. 19–36, mar. 2002. ISSN 0891-2017. Available from Internet: <https://doi.org/10.1162/089120102317341756>.

POWER, C. et al. Navigating, discovering and exploring the web: Strategies used by people with print disabilities on interactive websites. In: **Human-Computer Interaction – INTERACT 2013**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 667–684. ISBN 978-3-642-40483-2. Available from Internet: <https://link.springer.com/chapter/10.1007\%2F978-3-642-40483-2_47>.

RAYNER, K. et al. So much to read, so little time: How do we read, and can speed reading help? **Psychological Science in the Public Interest**, SAGE Publications, v. 17, n. 1, p. 4–34, jan. 2016. Available from Internet: <https://doi.org/10.1177/1529100615623267>.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2019. Available from Internet: <https://arxiv.org/abs/1908.10084>.

REIMERS, N.; GUREVYCH, I. Making monolingual sentence embeddings multilingual using knowledge distillation. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2020. Available from Internet: <https://arxiv.org/abs/2004.09813>.

REYNAR, J. **Topic segmentation: Algorithms and applications**. USA: University of Pennsylvania, 1998. Ph.D. Thesis, Computer and Information Science.

REYNAR, J. C. An automatic method of finding topic boundaries. In: **Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics**. USA: Association for Computational Linguistics, 1994. (ACL '94), p. 331–333. Available from Internet: <https://doi.org/10.3115/981732.981783>.

RIBEIRO, L. L. G.; FILHO, W. M. D. C.; LEITE, F. P. A. **Comentários Ao Estatuto Da Pessoa Com Deficiência: Lei N. 13.146/2015**. São Paulo: Saraiva, 2016. ISBN 8553606910.

SALTON, G. et al. Automatic text structuring and summarization. **Information Processing & Management**, v. 33, n. 2, p. 193 – 207, 1997. ISSN 0306-4573. Methods and Tools for the Automatic Construction of Hypertext. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0306457396000623>.

SOLBIATI, A. et al. Unsupervised topic segmentation of meetings with BERT embeddings. **arXiv preprint arXiv:2106.12978**, 2021. Available from Internet: <https://export.arxiv.org/abs/2106.12978>.

STRAKA, M. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In: **Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies**. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 197–207. Available from Internet: <https://www.aclweb.org/anthology/K18-2020>.

SUNDAR, S. S. Theorizing interactivity's effects. **The Information Society**, Routledge, v. 20, n. 5, p. 385–389, 2004. Available from Internet: <https://doi.org/10.1080/01972240490508072>.

SWELLER, J.; AYRES, P.; KALYUGA, S. Measuring cognitive load. In: **Cognitive Load Theory**. New York, NY: Springer New York, 2011. p. 71–85. ISBN 978-1-4419-8126-4. Available from Internet: <https://doi.org/10.1007/978-1-4419-8126-4_6>.

TOMBAUGH, J.; LICKORISH, A.; WRIGHT, P. Multi-window displays for readers of lengthy texts. **International Journal of Man-Machine Studies**, v. 26, n. 5, p. 597–615, 1987. ISSN 0020-7373. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0020737387800731>.

TRIPPAS, J. R. Spoken conversational search: Speech-only interactive information retrieval. In: **Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval**. New York, NY, USA: ACM, 2016. (CHIIR '16), p. 373–375. ISBN 978-1-4503-3751-9. Available from Internet: <http://doi.acm.org/10.1145/2854946.2854952>.

WANG, T.; REDMILES, D. Auditory overview of web pages for screen reader users. In: **Adjunct Publication of the 30th Annual ACM Symposium on User Interface Software and Technology**. New York, NY, USA: ACM, 2017. (UIST '17), p. 193–195. ISBN 978-1-4503-5419-6. Available from Internet: <http://doi.acm.org/10.1145/3131785.3131837>.

WARNICK, B. et al. Effects of Campaign-to-User and Text-Based Interactivity in Political Candidate Campaign Web sites. **Journal of Computer-Mediated Communication**, v. 10, n. 3, 07 2005. ISSN 1083-6101. JCMC10315. Available from Internet: <https://doi.org/10.1111/j.1083-6101.2005.tb00253.x>.

WEBAIM. **Screen Reader User Survey #8 Results**. 2019. Available from Internet: <https://webaim.org/projects/screenreadersurvey8/>.

WEBAIM. **Screen Reader User Survey #9 Results**. 2021. Available from Internet: <https://webaim.org/projects/screenreadersurvey9/>.

WIERINGA, R. et al. Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. **Requirements Engineering**, Springer Science and Business Media LLC, v. 11, n. 1, p. 102–107, nov. 2005. Available from Internet: <https://doi.org/10.1007/s00766-005-0021-6>.

World Health Organization; World Bank (Ed.). **World report on disability**. Geneva, Switzerland: World Health Organization, 2011. ISBN 9789241564182. Available from Internet: <https://www.who.int/publications/i/item/9789241564182>.

YARRINGTON, D.; MCCOY, K. Approaches to locating areas of interest related to questions in a document for non-visual readers. In: **Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility**. New York, NY, USA: ACM, 2009. (Assets '09), p. 225–226. ISBN 978-1-60558-558-1. Available from Internet: <http://doi.acm.org/10.1145/1639642.1639687>.

YE, M.; LI, P.; LI, Q. An intelligent mobile news reader for the visually impaired people. In: **Proceedings of PACIS 2014**. Chengdu, China: PACIS, 2014. p. 130. Available from Internet: <https://aisel.aisnet.org/pacis2014/130>.

YE, M.; LI, P.; LI, Q. Vipreader: A light news reader for the visually impaired person. In: **2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent**

Agent Technologies (IAT). IEEE, 2014. v. 2, p. 282–287. Available from Internet: <https://doi.org/10.1109/wi-iat.2014.109>.

APPENDIX A – Systematic mapping of the literature

This study had the objective to provide an overview of previous studies in skim-reading techniques with screen readers. The search string (Table 1) was composed of synonyms for each area of the study (visually impaired people, skimming and screen readers). The terms used were calibrated using keywords from selected studies to improve the comprehensiveness of the string. To validate the search strings, three papers were used as a control group: Ahmed et al. (2013), Gadde and Bolchini (2014) and Giraud, Thérouanne and Steiner (2018). The chosen research databases were Scopus, ACM, Web of Science and Engineering Village, for their popularity in computer engineering area, and wide range of journals and conferences.

Table 1 – Databases and Search Strings Used

Database	Search String
Scopus	(TITLE-ABS-KEY (((summari?ation) OR (topicali?ation) OR (skimming) OR (speed AND reading)) AND ((blind) OR (visually AND impaired) OR (visual AND impairment)) AND ((assistive AND technology) OR (screen AND reader) OR (software))))
ACM	recordAbstract:((summarization OR summarisation OR topicalization OR topicalisation OR skimming OR "speed reading") AND (blind OR "visually impaired" OR "visual impairment") AND ("assistive technology" OR "screen reader" OR software)
Web of Science	TI=((summari?ation OR topicali?ation OR skimming OR (speed AND reading)) AND (blind OR (visually AND impaired) OR (visual AND impairment)) AND ((assistive AND technology) OR (screen AND reader) OR software))
Engineering Village	((summari?ation) OR (topicali?ation) OR (skimming) OR (speed AND reading)) AND ((blind) OR (visually AND impaired) OR (visual AND impairment)) AND ((assistive AND technology) OR (screen AND reader) OR (software))

Source: from the author (2021)

To filter through searched papers before conducting the full-test reading, inclusion and exclusion criteria were applied based on the research questions.

Inclusion Criteria: The inclusion criteria were applied during the search stage, before accessing the full text, by using the title, keywords and abstract provided on the web.

IC1: The abstract explicitly mentions assistive technologies to ease the use of screen readers by visually impaired people.

IC2: From the abstract, the researcher can deduce that the focus of the paper contributes to assistive technologies to help screen reader users in information access.

IC3: Only studies with English abstracts and titles were included.

Exclusion Criteria: The exclusion criteria were applied after the initial database searching process. This is necessary to filter studies not related to the research questions that were included in the previous stage.

EC1: Duplicates were be excluded.

EC2: The study lies outside the software engineering domain.

EC3: Contributions are not related to visually impaired people, screen readers or skimming techniques, terms are only mentioned in the general introductory sentences of the abstract or are not further discussed.

EC4: Studies involving only Braille, not using screen readers.

The research databases returned 235 studies in total and after applying the inclusion and exclusion criteria, 20 studies remained. To further filter the selected studies, a full-text reading was done to exclude studies that did not contribute to the area and studies that were inconclusive. The final result was 16 studies.

Keywording of Abstracts

Following the systematic mapping process described by Petersen et al. (2008), we used the Keywording as a way to reduce the time needed to develop a classification scheme. It starts by reading abstracts to identify keywords and concepts reflecting the contribution of the paper. Then, the set of keywords from different papers are combined and clustered to form the categories for the map.

This study employed two main facets. The first facet was the approach used, for example, if the proposal provided a reading aid by information filtering, summarization or concurrent speech.

The second facet was the research type, which reflected the research approach used and was based on the classification proposed by Wieringa et al. (2005), summarized in Table 2.

Table 2 – Research Type Facet

Category	Description
Validation Research	Novel techniques that had not been yet been implemented in practice.
Evaluation Research	An evaluation was conducted on already implemented techniques.
Solution Proposal	A solution for a problem is proposed, being novel or a significant extension of an existing technique.
Philosophical Papers	Papers that structure a field in form of a taxonomy or a conceptual framework, sketching a new way of looking at existing things.
Opinion Papers	These papers express personal opinion on a technique or how things should be done.
Experience Papers	Papers that portray the personal experience of the author by explaining what and how something has been done.

Source: from the author (2021)

APPENDIX B – Texts used in the user study

Text 1 – Novo coronavírus infecta e se replica em células das glândulas salivares

Pesquisadores da Faculdade de Medicina da Universidade de São Paulo (FMUSP) constataram que o coronavírus infecta e se replica em células das glândulas salivares.

Por meio de análises de amostras de três tipos de glândulas salivares, obtidas durante um procedimento de autópsia minimamente invasiva em pacientes que morreram em decorrência de complicações da Covid-19 no Hospital das Clínicas da FMUSP, eles verificaram que esses tecidos especializados na produção e secreção de saliva são reservatórios para o novo coronavírus. Os resultados do estudo, apoiado pela Fapesp, foram publicados no Journal of Pathology.

As descobertas contribuem para explicar por que o novo coronavírus é encontrado em grandes quantidades na saliva, o que viabilizou a realização de testes para diagnósticos da Covid-19 a partir do fluido, sublinham os autores do trabalho.

"É o primeiro relato de vírus respiratório capaz de infectar e se replicar nas glândulas salivares. Até então, acreditava-se que apenas vírus causadores de doenças com prevalência muito alta, como o da herpes, usavam as glândulas salivares como reservatório. Isso pode ajudar a explicar por que o Sars-CoV-2 é tão infeccioso", diz à Agência Fapesp Bruno Fernandes Matuck, doutorando na Faculdade de Odontologia da USP e primeiro autor do estudo.

Os pesquisadores já tinham demonstrado, em estudo anterior, a presença de RNA do Sars-CoV-2 no tecido periodontal de pacientes que morreram em decorrência da Covid-19.

Em razão da alta infecciosidade do Sars-CoV-2 quando comparado a outros vírus respiratórios, eles levantaram a hipótese de que o novo coronavírus poderia infectar e se replicar em células das glândulas salivares e, dessa forma, surgir na saliva sem ter contato com secreções nasais e pulmonares.

Isso porque estudos internacionais anteriores mostraram que o ducto salivar apresenta o receptor ACE2, com o qual a proteína spike do Sars-CoV-2 se liga para infectar as células. Mais recentemente, outros grupos de cientistas relataram ter observado em estudos feitos com animais que, além da ACE2, receptores como a serina protease transmembranar 2 (TMPRSS) e a furina, presentes nos tecidos das glândulas salivares, são alvos do Sars-CoV-2.

A fim de testar essa hipótese em humanos, foram feitas biópsias guiadas por ultrassom em 24 pacientes que morreram em decorrência da Covid-19, com idade média de 53 anos, para extração de amostras de tecidos das glândulas parótida, submandibular e menores.

As amostras dos tecidos foram submetidas a análises moleculares (RT-PCR) para identificação da presença do vírus. Os resultados indicaram a presença do vírus nos tecidos em mais de dois terços das amostras.

Já por meio de marcações imuno-histoquímicas – em que é colocado um corante em uma molécula que se gruda no vírus e nos receptores –, foi possível observar a presença do vírus in situ, no interior dos tecidos. E, por meio de microscopia eletrônica, foi detectada não só a presença, mas também o vírus se replicando nas células e identificado o tipo de organela que ele utiliza para essa finalidade.

“Observamos vários vírus aglomerados nas células das glândulas salivares – um indicativo de que estão se replicando em seu interior. Não estavam presentes nessas células passivamente”, afirma Matuck.

Os pesquisadores pretendem avaliar, agora, se a boca pode ser uma porta de entrada direta do Sars-CoV-2, uma vez que os receptores ACE2 e o TMPRSS são encontrados em vários locais da cavidade, como em tecidos da gengiva e da mucosa bucal. Além disso, a boca tem uma área de contato maior do que a cavidade nasal, apontada como a principal porta de entrada do vírus.

“Por meio de uma parceria com pesquisadores da Universidade da Carolina do Norte, dos Estados Unidos, pretendemos mapear a distribuição desses receptores na boca e quantificar as replicações virais em tecidos bucais”, diz Luiz Fernando Ferraz da Silva, professor da FMUSP e coordenador do projeto. “Pode ser que a boca seja um meio viável para entrada direta do vírus”, estima Matuck.

Outra ideia é verificar se idosos possuem mais receptores ACE2 na boca em comparação com pessoas mais jovens, uma vez que têm uma diminuição do fluxo salivar. A despeito disso, os pesquisadores encontraram mesmo em pacientes idosos, que têm menos tecidos salivares, uma alta carga viral.

“Esses pacientes quase não tinham tecido salivar, era quase tudo tecido gorduroso. Mas, mesmo assim, ainda apresentavam uma carga viral relativamente alta”, relata Matuck.

Text 2 – Covid: Risco de infecção por coronavírus varia ’muito’ de acordo com máscara usada, diz estudo Usar uma máscara PFF3 de alto grau pode proteger quase inteiramente os profissionais de saúde da covid, constatou a pesquisa.

A qualidade das máscaras faciais que os profissionais de saúde usam faz uma grande diferença no risco de infecção por coronavírus, segundo um estudo conduzido em um hospital em Cambridge, no Reino Unido.

Usar uma máscara de alto grau conhecida como PFF3 pode fornecer até 100% de proteção.

Já as máscaras cirúrgicas comuns apresentam um risco muito maior de se pegar o vírus, segundo a pesquisa.

Entidades de classe vem fazendo campanha para que trabalhadores recebam equipamentos melhores de proteção individual (EPI).

Os dados foram coletados durante um programa de testes da entidade NHS Foundation Trust, ligada ao sistema de saúde britânico, o NHS (equivalente ao SUS no Brasil).

Os resultados foram publicados em um artigo que ainda não foi revisado por pares.

Durante a maior parte do ano passado, o hospital seguiu a orientação do governo que especifica que os profissionais de saúde devem usar máscaras cirúrgicas, exceto em algumas situações limitadas.

Embora resistentes a fluidos, essas máscaras são relativamente frágeis e frouxas e não têm o objetivo de filtrar aerossóis infecciosos — minúsculas partículas de vírus que podem permanecer no ar e agora são amplamente tidas como fonte de infecção por coronavírus.

O estudo descobriu que a equipe que cuida de pacientes em alas "vermelhas" (onde há pacientes com covid) enfrentava um risco até 47 vezes maior do que aqueles em alas "verdes" (sem pacientes com covid).

O pesquisador principal, Mark Ferris, especialista em saúde ocupacional do hospital, disse que a equipe estava se contaminando com covid, apesar de seguir todos os protocolos de controle de infecção.

Assim, quando a segunda onda da pandemia começou a acontecer em dezembro passado, os gerentes dos hospitais de Cambridge tomaram a decisão de aumentar a proteção nas alas vermelhas.

"A única coisa que faltava tentar fazer ainda era adotar as máscaras PFF3, e foi isso que eles fizeram", disse Ferris.

As máscaras PFF3 são mais fechadas no rosto e projetadas especificamente para filtrar aerossóis.

Nas semanas seguintes a essa mudança, a taxa de infecções entre os profissionais de saúde nas enfermarias vermelhas caiu drasticamente, atingindo o nível das alas verdes, onde não havia pacientes com covid.

O estudo conclui que "os casos atribuídos à exposição na ala caíram significativamente, com máscaras PFF3 fornecendo 31-100% de proteção (e provavelmente 100%) contra a infecção de pacientes com covid-19."

Todos os casos restantes provavelmente foram causados por disseminação na comunidade, e não no hospital.

O artigo afirma que as máscaras cirúrgicas resistentes a fluidos são "insuficientes" para proteger os profissionais de saúde.

Mike Weekes, da Cambridge University NHS Hospitals Foundation Trust, que trabalhou no estudo, disse que a pesquisa dá "algumas evidências do mundo real de que as máscaras PFF3 são realmente eficazes, e mais eficazes do que as máscaras cirúrgicas".

Ele acrescentou: "Claramente, é um estudo relativamente pequeno e, portanto, precisamos ver essas descobertas replicadas em outro lugar. Mas, dada a diferença nos resultados que vimos, como uma espécie de efeito do princípio de precaução, o que devemos pensar é adotar máscaras PFF3 para quem cuida de um paciente com coronavírus."

O hospital de Cambridge está entre 17 em todo o Reino Unido que decidiram atualizar suas máscaras, à revelia da política nacional.

O apelo para que as máscaras PFF3 sejam distribuídas de forma mais ampla é uma demanda de longa data de diversos órgãos de profissionais de saúde.

Em uma carta aberta ao Secretário de Saúde do Reino Unido, Sajid Javid, um grupo de consultores e médicos diz que o novo estudo fornece ainda mais evidências de por que a política precisa ser alterada.

"Isso tem implicações importantes para a proteção do trabalhador de saúde, já que o Reino Unido lida com o que se espera que seja uma 'onda final', além de tentar reduzir o enorme acúmulo de outros trabalhos enquanto enfrenta a inevitável doença e isolamento da equipe", escreveu o grupo.

Um porta-voz do Departamento de Saúde e Assistência Social disse que as orientações sobre os padrões de máscaras são atualizadas regularmente para refletir os mais recentes avanços científicos.

"A segurança do NHS (sistema de saúde britânico) e da equipe de assistência social sempre foi nossa prioridade e continuamos a trabalhar 24 horas por dia para entregar materiais para proteger aqueles na linha de frente. As evidências e os dados emergentes são monitorados e revisados continuamente e as orientações serão alteradas de acordo, caso seja apropriado."

Text 3 – Por que a vida dos seus amigos parece ser melhor do que a sua

Você já sentiu como se todo mundo tivesse mais a agradecer do que você? Cheque o seu feed o Facebook ou do Instagram: seus amigos parecem jantar mais em restaurantes chiques, tirar mais férias exóticas e ter filhos mais bem-sucedidos. Eles têm até bichos de estimação mais fofos!

Pode ter certeza de que isso é uma ilusão, que está enraizada em uma propriedade das relações sociais conhecida como o paradoxo da amizade. Inicialmente formulado pelo sociólogo Scott Feld, o paradoxo afirma que “seus amigos são, em média, mais populares do que você”. Essa propriedade se combina a outras peculiaridades das interações sociais para criar uma ilusão.

O que o paradoxo significa é: se eu te perguntasse quem são os seus amigos e então os conhecesse, de forma geral eu consideraria que eles têm mais conexões sociais do que você. É claro, se você é uma pessoa extremamente sociável, o paradoxo não se aplicará a você. Mas para a maioria de nós, é provável que ele se aplique.

Embora esse paradoxo possa ocorrer em qualquer relação social, isso acontece de forma desenfreada online. Um estudo concluiu que 98% dos usuários do Twitter seguem contas que têm mais seguidores do que eles.

Embora soe estranho, o paradoxo da amizade tem uma explicação matemática simples.

O círculo pessoal de cada pessoa é diferente. A maioria de nós tem alguns amigos, e aí existem as pessoas bem conectadas, como David Rockefeller, antigo CEO do Chase Manhattan Bank, cuja agenda inclui mais de 100 mil pessoas.

Nas redes sociais, celebridades como Justin Bieber podem ter mais de 100 milhões de seguidores. É esse pequeno grupo de pessoas superconectadas — pessoas com muitos amigos, que são parte do seu círculo social — que aumentam a popularidade média dos seus amigos.

Esse é o truque matemático no centro do paradoxo da amizade. A popularidade extraordinária da pessoas como Justin Bieber não só desvirtua a popularidade média de amigos para qualquer um que esteja conectado a eles, mas mesmo que pessoas como ele sejam raras, elas também aparecem em um número extraordinário de círculos sociais.

E o paradoxo da amizade não é apenas uma mera curiosidade matemática. Ele tem aplicações úteis em prever tendências e monitorar doenças. Pesquisadores o utilizaram para prever temas em alta no Twitter semanas antes de eles acontecerem e para identificar surtos de gripe em suas fases iniciais e elaborar estratégias eficientes para gerenciar a doença.

Funciona assim: imagine, por exemplo, que você chega a uma vila africana com apenas cinco doses de vacina contra o Ebola. A melhor estratégia não é vacinar as primeiras cinco pessoas que você encontrar, mas sim perguntar a essas pessoa quem são os seus amigos e vacinar esses cinco amigos. Se você fizer isso, é mais provável que você escolha pessoas que têm círculos sociais mais amplos e assim poderiam infectar mais pessoas se ficassem doentes. Vacinar os amigos seria mais eficiente para conter a disseminação do Ebola do que imunizar pessoas aleatórias que podem estar na periferia de um círculo social e não conectadas a muitas outras pessoas.

Tem mais. Notavelmente, uma versão mais forte do paradoxo da amizade se aplica a muitas pessoas. A maior parte dos seus amigos têm mais amigos do que você. Pense nisso. Não estou mais falando sobre médias, onde um amigo excepcionalmente popular poderia desvirtuar a média de popularidade dos seus amigos.

O que isso significa é que a maioria dos seus amigos é mais bem conectada socialmente do que você. Vá em frente e experimente. Clique no nome de cada um dos seus amigos no Twitter e veja quantos seguidores eles têm e quantas contas eles seguem. Estou disposta a apostar que os números são maiores do que os seus.

Ainda mais estranho, esse paradoxo se aplica não só à popularidade, mas também a outras características, como o entusiasmo ao usar redes sociais, jantares em bons restaurantes, ou tirar férias em lugares exóticos. Como um exemplo concreto, veja com que frequência alguém posta atualizações no Twitter.

É verdade que a maioria das pessoas que você segue posta mais atualizações do que você. Além disso, a maioria das pessoas que você segue recebe mais informações novas e diversas do que você. E a maioria das pessoas que você segue recebe mais informações virais que acabam se espalhando no seu feed mais do que você imagina.

Essa versão mais forte do paradoxo da amizade pode levar a uma ilusão de maioria, na qual uma característica que é rara em uma rede de forma geral parece comum em muitos círculos sociais. Imagine que poucas pessoas, no geral, são ruivas, mas ainda assim, para muitas pessoas, parece que a maioria dos seus amigos tem cabelo ruivo. Tudo o que precisamos para disseminar a ilusão de que o cabelo ruivo é comum é que alguns influenciadores superconectados sejam ruivos.

A ilusão da maioria pode explicar por que você pode achar que seus amigos parecem estar fazendo coisas mais empolgantes: pessoas que são mais conectadas socialmente influenciam desproporcionalmente o que vemos e aprendemos nas redes sociais. Isso ajuda a explicar por que adolescentes superestimam a prevalência de bebedeiras em campi de universidades e por que alguns temas parecem ser mais populares no Twitter do que realmente são.

A ilusão da maioria pode distorcer as suas percepções sobre a vida de outros. Pessoas mais bem conectadas também fazem coisas mais notáveis, como jantar em um restaurante premiado ou tirar férias em Bora Bora. Elas também são mais ativas em redes sociais e mais propensas a registrar a vida no Instagram, distorcendo as nossas percepções de quão comuns essas coisas realmente são. Um bom jeito de mitigar essa ilusão é parar de se comparar com amigos e ser grato pelo que você tem.

Text 4 – Largada para o 2º Enem da pandemia

Começam hoje as inscrições para a edição 2021 do Exame Nacional do Ensino Médio (Enem), abertas até o 14 de julho. As provas estão marcadas para os dias 21 e 28 de novembro, nesta que será a segunda edição do Enem

desde o início da pandemia de COVID-19. Especialistas avaliam que não haverá grande mudanças no que se refere ao conteúdo e apostam na redução da tensão em relação à última edição, com estudantes agora mais acostumados ao ensino a distância e até mesmo já podendo acessar as aulas presenciais nos cursinhos preparatórios. Lembram, no entanto, que essa adaptação é desigual, principalmente para os pobres. Um calendário mais definido também contribui para uma expectativa melhor, na visão de alunos que enfrentaram as incertezas na versão passada do Enem, realizada este ano depois de adiamentos.

Para se inscrever, os candidatos devem acessar a página do participante e escolher entre fazer a prova impressa ou digital. Este ano, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), responsável pela realização do Enem, vai oferecer mais de 101 mil inscrições para a versão digital da prova. Apenas alunos que já concluíram o ensino médio ou vão concluir em 2021 podem optar por esse modelo. As provas impressa e digital serão feitas na mesma data e, portanto, terão as mesmas questões e proposta de redação.

Essa é a segunda edição do exame realizada durante a pandemia. “Para este ano, não esperamos grandes mudanças. É fato que todo ano tem algumas pequenas variações. Às vezes uma prova mais fácil, outra mais difícil, mas não é uma mudança drástica. Acreditamos que o Enem não deva ter mudanças estruturais”, avalia o diretor de ensino do Grupo Bernoulli, Rommel Domingos.

Para Domingos, entretanto, parte dos alunos pode enfrentar mais dificuldades devido à preparação de forma remota. “A preparação do aluno foi afetada desde o ano passado, em função da pandemia. Grande parte deles se adaptou bem ao ensino remoto, conseguiu desenvolver e crescer apesar das dificuldades. Está agora se preparando para o Enem e não deve ter perda no seu rendimento. Mas é fato que há um grupo de alunos que não se adaptou e está tendo dificuldade de administrar o seu tempo e estudar. Em particular, os alunos mais pobres, que têm menos espaços em sua residência, não têm bons equipamentos nema internet de qualidade."

Outro que não acredita em mudança no nível das provas é o diretor da Associação Pré-PUC e UFMG Pré-Vestibular, Richard Thuin. "A maior alteração seria a prova on-line. As duas vão ser no mesmo dia este ano. Quanto ao conteúdo não vejo diferença não." Ele também destaca que a pandemia dificultou a preparação dos alunos, principalmente os mais carentes.

"Eles tiveram aula on-line, alguns conseguiram acompanhar, outros não. Em função disso, estamos formando muitos alunos no EJA (Educação de Jovens e Adultos). Principalmente alunos de 3° ano que não concluíram (os estudos) na escola pública e estão buscando o curso EJA. Com a pandemia, os pais de alguns alunos perderam o emprego e os filhos precisaram buscar atividades para ajudar no sustento da casa. Com isso, alguns pararam de estudar e resolveram terminar o curso com o EJA, por ser mais rápido. E ele habilita o aluno a fazer as provas do Enem", explica.

Mas, para Rommel Domingos, existe um fator positivo para a edição deste ano. "A ansiedade é menor. Este ano a pandemia continua. O retorno do pré-vestibular já aconteceu e do colégio ainda não, estamos aguardando. É um retorno muito comedido, muitos alunos ainda continuam em casa. Mas a ansiedade este ano é muito menor que a do ano passado. O ambiente não está tão pesado", acredita.

Para ele, o fato dos cursinhos já terem retomado as aulas presenciais pode ajudar nesse problema. "Tem a questão do pré-vestibular já ter voltado. Então, vêm para a aula presencial justamente os que precisam mais. O aluno que já está mais adaptado, continua estudando em casa. E o aluno que está sofrendo mais e não conseguiu se adaptar, acaba entrando no rodízio e sendo beneficiado. Foi importantíssima a volta ao presencial, nesse caso."

Richard Thuin acredita que as aulas on-line exigem um esforço maior dos alunos. "Na aula on-line, os alunos ficam mais tímidos e não perguntam. As aulas ficam gravadas. O curso remoto precisa de muita dedicação. Com as vídeoaulas, apesar de o aluno ter mais carga de conhecimento e mais tempo, é mais superficial. O aluno não aprende o contexto todo. Esse aluno vai precisar de mais interesse e, alguns estudantes nessa idade têm pouco interesse." Apesar disso, ele afirma que a aprovação a última edição do Enem não foi afetada. "Em 2020, diversos alunos de escolas públicas continuaram as aulas mesmo a distância e tivemos uma aprovação muito expressiva, de 86%."

Ana Beatriz Gontijo tem 19 anos e vai fazer o Enem pela segunda vez. Ela se formou em 2020 e pretende estudar medicina. No último Enem, as incertezas sobre o processo a deixaram angustiada. "No início deste ano, foi a primeira vez que fiz (o Enem) valendo, não como treineira. No início, a adaptação às aulas on-line foi muito difícil. E a confusão toda de fazer enquete, disseram que seria uma data, depois foi em outra. Só foi adiando. Isso foi muito angustiante, a gente não tinha nenhum cronograma para seguir e planejar".

Ela acredita que o impasse sobre a realização do exame pode ter atrapalhado seu desempenho. "Não posso colocar a culpa 100% nisso, mas acho que prejudicou sim. Foi uma prova muito diferente em relação ao conteúdo. Por exemplo, história do Brasil, não caiu praticamente nenhuma questão. E os professores nos prepararam de acordo com as experiências anteriores do Enem."

A estudante conta que, este ano, a preparação está sendo mais tranquila. "Como a gente já teve a experiência do on-line de 2020, agora está sendo um pouco mais fácil. Já peguei o ritmo, o jeito de estudar em casa. Inclusive, estou no pré-vestibular em que as aulas presenciais já voltaram (no modelo híbrido), mas nem estou indo porque já me acostumei à rotina de estudar em casa."

Outro ponto positivo, segundo Ana Beatriz, é o exame já ter data definida. No ano passado, a incerteza persistiu até o fim, inclusive com medidas judiciais para suspender o exame nacional.

Já Wanderson dos Santos, de 24 anos, vive uma situação diferente. Ele vai tentar o Enem pela terceira vez, sempre estudou em escola pública municipal e terminou o ensino médio em 2015. Também quer estudar medicina, mas não conseguiu se adaptar ao ensino remoto. "Foi uma experiência muito ruim por causa da distração. Na aula presencial, a gente tem um compromisso maior. Na aula on-line não, como ela é gravada, você começa a fazer outras coisas e quando vê está com um monte de matéria acumulada".