Dynamik: Syntactically-Driven Dynamic Font Sizing for Emphasis of Key Information

Naoto Nishida
The University of Tokyo
Bunkyo, Tokyo, Japan
nawta@g.ecc.u-tokyo.ac.jp

Yoshio Ishiguro
The University of Tokyo
Bunkyo, Tokyo, Japan
ishiy@acm.org

Jun Rekimoto
The University of Tokyo
Bunkyo, Tokyo, Japan
Sony CSL Kyoto

Naomi Yamashita
Social Informatics
Kyoto University
Sakyo, Kyoto, Japan

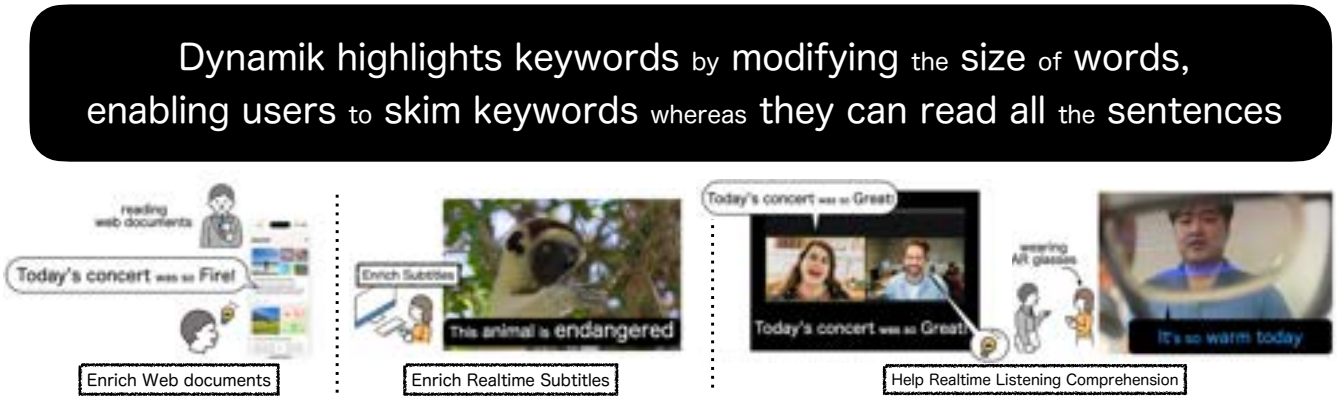


Figure 1: Dynamik is an automated keyword highlighting system for non-native speakers in a specific language when they use audio speech recognition to aid their listening skill. it highlights important keywords by modifying the size of words, enabling users to skim keywords whereas they can read all the sentences.

Abstract

In today's globalized world, there are increasing opportunities for individuals to communicate using a common non-native language (lingua franca). Non-native speakers often have opportunities to listen to foreign languages, but may not comprehend them as fully as native speakers do. To aid real-time comprehension, live transcription of subtitles is frequently used in everyday life (e.g., during Zoom conversations, watching YouTube videos, or on social networking sites). However, simultaneously reading subtitles while listening can increase cognitive load.

In this study, we propose Dynamik, a system that reduces cognitive load during reading by decreasing the size of less important words and enlarging important ones, thereby enhancing sentence contrast. Our results indicate that Dynamik can reduce certain aspects of cognitive load, specifically, participants' perceived performance and effort among individuals with low proficiency in English, as well as enhance the users' sense of comprehension, especially among people with low English ability. We further discuss our methods' applicability to other languages and potential improvements and further research directions.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**.

Keywords

Listening, Subtitling, Skimming, Keyword extraction

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
IUI '25, Cagliari, Italy
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712115

ACM Reference Format:
Naoto Nishida, Yoshio Ishiguro, Jun Rekimoto, and Naomi Yamashita. 2025. Dynamik: Syntactically-Driven Dynamic Font Sizing for Emphasis of Key Information. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy*. ACM, New York, NY, USA, 30 pages. https://doi.org/10.1145/3708359.3712115

1 Introduction

In today's globalized society, there are more and more situations in which people speak common language (*i.e.*, lingua franca) and communicate with each other. However, for many people around the world, speaking a language that is not their native tongue can cause difficulties, especially when they listen to native speakers [102]. To address this problem, subtitling has become widely used in videoconferencing systems and online video platforms as a listening aid [1, 37]. However, listening to audio and reading subtitles at the same time involves performing different tasks simultaneously, further increasing the cognitive load of non-native speakers [49, 61, 73]. Thus, there is a demand for subtitles that allow users to quickly obtain information on important parts of a text, rather than conventional subtitles [15, 42].

In past studies, many psychological experiments have been conducted to investigate the effects of various styles of texts (*e.g.*, fonts, sizes, etc.) on reader psychology, but there are limited studies that focus on alleviation of cognitive load. As examples of limited studies, Pan et al. showed that it decreases the burden on non-native users to have native speakers highlight the important parts of the subtitles by touching each word to allow non-native speakers to see the highlighted subtitles when they listen to native speakers [74]. Since native speakers have more leeway to annotate important keywords compared to non-native speakers, native speakers annotate keywords in this research. However, the automation of keyword would benefit both because native speakers can focus solely on the topic they are speaking on, and the other behaviors such as gestures or observation of the listeners. Hausataari et al. examined whether terminology-highlighted subtitles alleviate nonnative speakers' workload necessary for reading, in the case of catching up the conferences' speech [41]. The authors proved that termiology-highlighted subtitles did not improve the comprehension of nonnative speakers, but since speech recognition accuracy was not so good in that period (Word Error Rate was 23%), it is assumed that terminology could not be well recognized at the time.

In this study, we propose a system called "*Dynamik*" to address the alleviation of users' workload by automation of keyword highlighting in real-time. *Dynamik* classifies the importance of words in speech-recognized sentences into content words and function words based on morphological analysis and displays less important words (*i.e.*, function words) smaller and more important words (*i.e.*, content words) larger in real time, thus displaying the contrast of importance of each word visually to improve readability (Figure. 1). This method is expected to reduce the workload of non-native speakers listening and to achieve more effective information transfer. We conducted a crowd-sourcing experiment with 84 participants to investigate the effectiveness of our method in English.

The result showed that *Dynamik* was significantly preferred by non-native English speakers but not by native English speakers, that the workload items of *Performance* and *Effort* were significantly lighter and increased the sense of comprehension than under the other conditions such as common subtitles and the subtitles that only show keywords.

Our method is applicable to languages other than English, and we expect to reduce the display area of subtitles by reducing the size of unimportant words.

Our contributions are presented below:

- We investigated the possible automation design of accessible subtitles for non-native English speakers that works in realtime. To the best of our knowledge, this is the first work to automate the highlighted keyword subtitle from morphological aspects.
- We developed and published a real-time dynamic subtitle system which we applied to build *Dynamik* [1] [2]. This system can be used for other methods to modify subtitles with subtle changes (*e.g.*, changing font color, other definition of keywords other than morphoanalysis).
- We presented *Dynamik*, a novel subtitling method to assist non-native English speakers during listening, focusing on function words and content words. No prior work used these perspectives for workload alleviation.
- We published a web app example of secure psychological experiments for crowd-sourcing by encryption of external files and obfuscation of codes, which we used for our experiment [3]. This code is especially useful for when you have something you don't want participants to read in the projects (in our case, the answer of quizzes and completion codes).
- We conducted evaluation experiments with 84 people and showed that *Dynamik* alleviates several workloads (*Effort*, *Performance*) and affords more self-awareness of comprehension for non-native English speakers.

2 Related Work

Our research focuses on alleviating workloads during non-native English speakers' communication. First, we discuss a communication challenge for non-native English speakers and existing solutions for it. Second, we discuss existing research methods on media richness for subtitles to deepen our solution. Third, we discuss several examples for extracting keywords or key phrases to discuss our concrete implementation. Lastly, we further discuss possible applications of our research to consider our keyword extraction method.

2.1 Communication Challenge for Non-Native English Speakers

Non-native English speakers often find themselves compelled to communicate in English, even though they do not have the same level of comprehension as native English speakers. This situation creates a significant cognitive load that can be further exacerbated by the provision of full, unfiltered subtitles. A potential solution to this problem might involve selective presentation of information, omitting less crucial elements, or highlighting important elements.

Although direct solutions to these issues have not been extensively proposed, related research in the fields of psychology, computer graphics, and computer-aided language learning (CALL) has produced various relevant findings [26, 44, 57, 97].

In the field of CALL, numerous studies have explored the use of audio and text subtitles for the learning of foreign languages. For

[1] https://github.com/nawta/Dynamik_client
[2] https://github.com/nawta/Dynamik_server
[3] https://github.com/nawta/Dynamik_experiment

example, Hwang compared the effects of fragmented subtitles versus standard subtitles on English learning outcomes and cognitive load [37]. The study concluded that fragmented subtitles increased cognitive load and were more effective in English learning. Another notable approach is *Flash Word*, which provides fragmented audio synchronized subtitles, using keywords extracted through tf-idf to assist English as a Second Language (ESL) learners [47, 111].

It is important to note that while these studies aim to increase cognitive load for learning purposes, our research focuses on reducing the cognitive load associated with reading subtitles during listening tasks.

2.2 Subtitle Display Methods

Research on enhancing lean media with rich elements has a long history in the field of media richness and accessibility [19, 43, 60, 87].

Regarding subtitling, many studies, particularly those focusing on Deaf and Hard of Hearing (DHH) individuals, have explored ways to incorporate non-verbal elements of speech into subtitles [6, 20, 54]. For individuals with dyslexia, some proposed ways to facilitate text skimming [80, 98]. These studies have examined various aspects of text presentation, including font size [29, 106], height[51], highlighting [48], font color [8, 75], typeface change [53], font weight [86, 98], appending emojis [55, 69], upper case [7], underlining [103], bold or italic [7, 86], transparency [63], text spacing [65] , syllables [91], removal of text [101], dynamic positioning [45, 71]. In addition, research in cognitive psychology has investigated how font size and other typographic elements influence text perception [14, 58, 109].

In the field of text design, principles have been established showing that variation in the emphasis of text can improve the comprehension of ideas [46]. This principle suggests that differentiating between keywords and non-keywords through visual emphasis could promote understanding.

Our research targets non-native speakers, and we have drawn inspiration from these previous studies on subtitle variations to create a design inspiration for our implementation.

2.3 Keyword Extraction Method from Linguistic Perspective

Several approaches to keyword extraction have been proposed in previous research in the field of Natural Language Processing, including statistical methods (derived from tf-idf) [16, 25, 47, 59, 88, 89, 112], graph-based methods (derived from TextRank) [11, 12, 28, 62, 96, 105], and some Seq2seq neural network models (derived from word embedding) [35, 90, 93]. However, most of these are supposed to be used a posteriori to preexisting transcripts. Seq2seq models can predict a priori, but still require training. To measure our Proof-of-Concept quickly, we first looked for a method that could be done without training.

From a linguistic perspective, content words (nouns, main verbs, adjectives, adverbs) carry specific meanings, while function words serve primarily grammatical roles and contribute less to the content of the sentence [17, 30]. In English speech, content words are typically stressed, while function words are unstrained, reflecting their relative contribution to information content [50].

Based on the aforementioned design principles and these findings, if content words are defined as keywords, it may be possible to promote comprehension by making content words more prominent in sentences while making function words less prominent.

Therefore, *Dynamik*, the system we propose in this study, integrates these findings by displaying function words in smaller text and content words in larger text, aiming to support non-native English speakers' listening comprehension.

2.4 Subtitle Display Area

Research on subtitle display areas and placement has been conducted primarily in the context of visualization studies [32, 110]. Minimizing the subtitle display area has been a goal in these studies, with potential benefits for devices with limited display areas, such as smart glasses.

The proportion of function words to content words in a text is referred to as Lexical Density, defined by the following formula by Halliday [39]:

$$\text{Lexical Density} = \frac{\text{Number of Content Words}}{\text{Total Number of Clauses}} * 100$$

Typically, function words account for approximately 40% of the words in a text [52]. By reducing the size of or omitting function words, we can potentially decrease the overall subtitle display area.

As discussed in our design section later, we chose to visually de-emphasize function words and emphasize content words to support non-native English speakers' listening comprehension while potentially reducing the subtitle display area.

3 Design

3.1 Core Concept

The central idea of this research is to enhance non-native English listeners' ability to efficiently extract information from text by emphasizing keywords and deemphasizing less crucial words. In short, our goal is to develop subtitles that facilitate more effective skimming.

3.2 Defining Keywords

The definition of content words (*i.e., words that possess semantic content and contribute to the meaning of the sentence in which they occur*) and function words (*i.e., words that have little lexical meaning or ambiguous meaning and express grammatical relationships among other words within a sentence*) has shown that the classification of words into content words and function words strongly correlates with their contribution to the overall meaning of a text [30]. As textual content positively correlates with information density, we can broadly categorize content words as essential and function words as less critical for our purposes. In addition, in phonology, the pronunciation patterns of words in speech can also inform this classification (*i.e.*, importance). Generally, function words are pronounced less prominently than content words [38]. Taking these factors into account, we have designated the following parts-of-speech as important "keyword":

- Nouns
- Verbs

- Adjectives
- Adverbs
- Negatives

3.3 Subtitle Presentation Method

Drawing from text design principles and variations in prior research on subtitles, we held a one-hour discussion session on several potential approaches for presenting assistive subtitles. The authors and one invited Natural Language Processing researcher participated in this session.

Regarding the color used in the experiment, we chose bright pink ((R, G, B) = (225, 128, 130)) for the texts and black for the background of Color Universal Design [100] because the previous study found that the combination of creme color and black is easy to read [82] and the dark black background is often used for AR systems.

Considering the potential applications of our system, we envisioned its use not only on computer monitors but also on devices where screen real estate is at a premium, such as Head-Mounted Displays (HMDs) and emerging smart glasses. In addition, there are options that are not suitable for all fonts (*e.g.*, bold, italic), leading to our method's lack of generalizability. Given this context, we prioritized minimization of screen occupancy in our design approach. Consequently, we chose to implement two methods to compare with normal subtitles (hereafter '**Normal**') (See Figure 2):

- Reducing the font size of function words (hereafter '**Dynamik**')
- Omitting function words entirely (hereafter '**Keyword**')

We named *Dynamik* after the term for musical expression through changes or contrasts in the intensity of the sound.

4 Implementation

We developed three types of assistive subtitles for our study: *Dynamik*, which reduces the size of less important words; Keyword, which completely omits less important words entirely; and *Normal*, standard subtitles as a control condition.

The system workflow is shown on Figure 3. It begins with capturing English audio input using the PC's built–in microphone. The audio is then processed through speech recognition, followed by morphological analysis of the real-time speech recognition results. Based on this analysis, words that are not considered content words (nouns, adjectives, verbs, and auxiliary verbs) are reduced in size or omitted, depending on the subtitle type. The processed text is then displayed on a Unity-based interface.

For the client-side implementation, we used Unity version 2022.3.21f1 with C# 12.0 and .NET Framework 8.0 [18, 22, 99], while the server-side was implemented in Python 3.10 [77]. We used the Azure Speech Recognition API for speech recognition [4] and spaCy 3.7.5 with en_core_web_sm 3.7.1 for morphological analysis and part-speech tagging [27, 95]. Part-of-Speech tagging was performed using a Hidden Markov Model implemented in spaCy, which also provided functionality for stop word detection. Communication between the client and the local server was facilitated using ZeroMQ [114], and between the client and the Azure server was facilitated using the Azure SDK for .NET [5]. The codes are available here [1] and here [2]. Our development and testing were conducted on a 13–inch MacBook Pro (2021 M1 model).

The system refreshes the subtitles every 0.5 seconds. Morphological analysis and parts of speech take approximately 0.2–0.3 seconds for English text (< 0.5 seconds) on the local server, and the communication between the client and the servers took infinitesimal compared to language processing. Therefore, we did not find significant differences in the subtitle presentation intervals between the different methods due to consistent processing times.

For the subtitle size, we used a standard font size of 18 pt for every condition, except for the case that *Dynamik*'s function words were displayed at 12 pt. These sizes were chosen according to the Web Content Accessibility Guidelines [108]. The guidelines says that "*with at least 18 point or 14 point bold or font size that would yield equivalent size for Chinese, Japanese, and Korean (CJK) fonts*" for large scale, and "*For many mainstream body text fonts, 14 and 18 points are roughly equivalent to 1.2 and 1.5 em or to 120 % or 150 % of the default size for body text*" for font jumps. The sizes are also based on the readability research on dyslexia, which about 10 % of people have) [64, 94]. The 18 pt size offers optimal readability for a broad audience, including those with dyslexia and the elderly [9, 84, 85], while the 12 pt is the smallest recommended size that maintains readability for the same diverse group [10, 83].

For subtitle typefaces, we chose ZenMaruGothic Medium [113], because it satisfies "monospaced, San Selif" features, which are supposed to be easy to read for both people with and without dyslexia [78, 79]. The color of the subtitles was selected as (R,G,B,A) = (255, 128, 130, 255), which is selected from a color palette for visually impared people [100].

5 Evaluation

To determine which assistive subtitle method offers the highest readability and reduces cognitive load, we conducted a crowd-sourced experiment involving 104 participants. This section details the participants, the experimental application, the procedure, and its result.

5.1 Participants

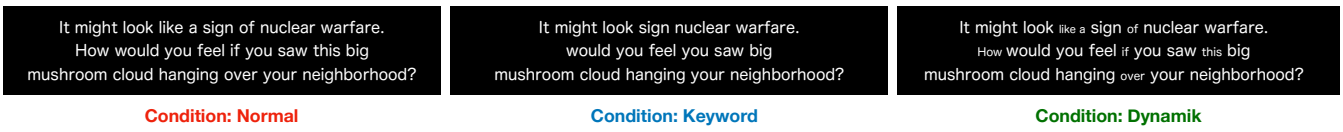
We used Prolific for crowd-sourcing [76]. Out of 104 initial participants, we excluded two who failed our dummy quiz, one who couldn't complete the experiment within the time limit (67 minutes, as calculated by Prolific), and 17 who didn't finish all responses. This left us with data from 84 participants for analysis. We included 18 native English speakers as a comparison group. We offer a £9.9 / hour – £12 / hour reward to the participants.

5.2 Experimental Application

We developed a custom experimental application using Vue.js 3.2.13 [104], Node.js v20.17.0 (npm v10.8.2) [67, 68], and Webpack 5.3.10 [33]. This setup allowed us to modify the directory structures between development and deployment, improving security. We used jsPsych 7.3.4 to create the experimental workflow [21] and Webpack Obfuscator 3.5.1 for the obfuscation of code [33]. The code is available here [3].

For data storage, we used jsPsychSheet [36], DataPipe [56], and the Open Science Framework [72].

For some figures inserted during the experiment for listening comprehension quizzes, we used Adobe Firefly [2].



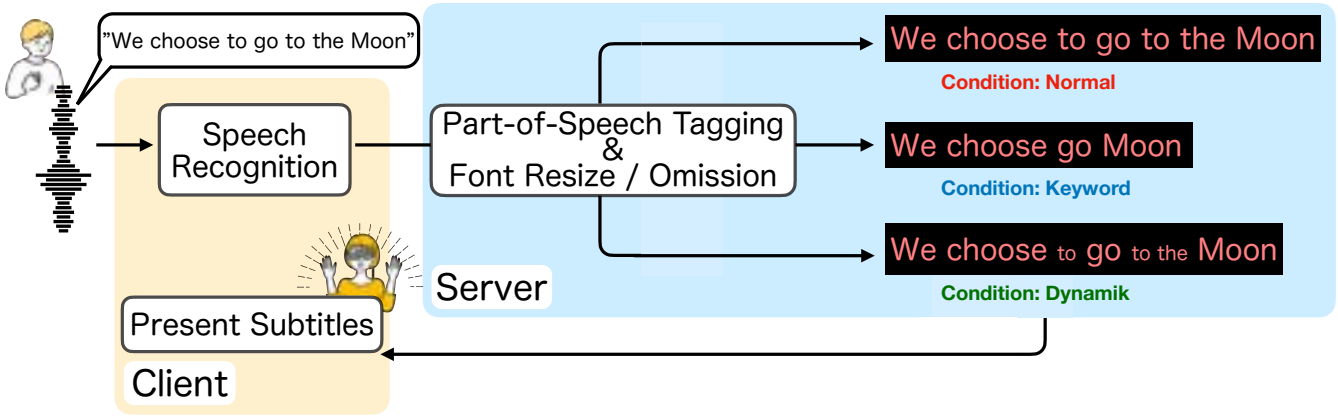


Figure 3: System workflow. It begins with capturing English audio input using the PC's built–in microphone. The audio is then processed through speech recognition, followed by morphological analysis of the real-time speech recognition results. Based on this analysis, words that are not considered content words (nouns, adjectives, verbs, and auxiliary verbs) are reduced in size or omitted, depending on the subtitle type. The processed text is then displayed on a Unity-based interface.

To protect highly credential content, such as question texts and completion codes, we encrypted external files using a combination of the XOR cipher, string concatenation/splitting, and Caesar cipher.

5.3 Procedure

The main part of the experiment workflow is shown in Figure 4. The detailed experimental workflow was as follows:

(1) Participants accessed the experiment page through Prolific.
(2) They consented to the experiment and provided basic information (gender, nationality, native language, other spoken languages, and English test scores such as TOEFL [23], TOEIC [24], IELTS [13]).
(3) Participants listened to an audio track and completed 10 TOEFL-adapted listening comprehension questions (hereafter '**Pre-test**') to assess their English listening skills. The audio track and five questions were from the 5th Edition TOEFL Official Guide (track 1) [92], and five were generated using Claude 3.5 Sonnet and validated by the authors (*e.g.*, so that the questions do not interfere with the others as hints) [3].
(4) The participants then listened to six CNN news clips [107] with assistive subtitles (See Table 2). We used three subtitle conditions (*Normal*, *Keyword*, and *Dynamik*), with two excerpts per condition presented in random order.
(5) After each excerpt, participants completed three questionnaires to ask their self-awareness of the extent of engagement with watching the clip, their self-awareness of the extent of comprehension of the clip, and the readability of the subtitle during the clip. After that they also completed NASA–TLX assessments and listening comprehension quizzes (hereafter '**Comprehension Quiz**') on the clip [40]. We incorporated dummy tests with a probability of appearance 40% on each NASA–TLX and the *Comprehension Quiz* page.
(6) After repeating step (5) six times, the participants commented on their impressions of the experiment to conclude the experiment.

The order of the answer choices for *Pre-test* and *Comprehension Quiz* was randomized. We adjusted rewards based on *Comprehension Quiz* performance, offering £3.3 for standard completion (estimated at 20 minutes, £9.9/hour) and £4 for scores above 80% (£12 / hour), although no participant achieved this threshold. The NASA–TLX and *Comprehension Quiz* are given six times, which means that the probability of no dummy quiz on each item is $(6/10)^6 = 0.046\ldots < 0.05$. The CNN video clips were 40–60 seconds long, with a resolution of 960 × 540, a video quality of 21 Mbps and an audio quality of 48.0 kHz. All kinds of subtitles were updated on the screen every 0.5 seconds. The average duration of the experiment was 25 minutes and 47 seconds. We also had a free-form questionnaire at the end of the experiment.

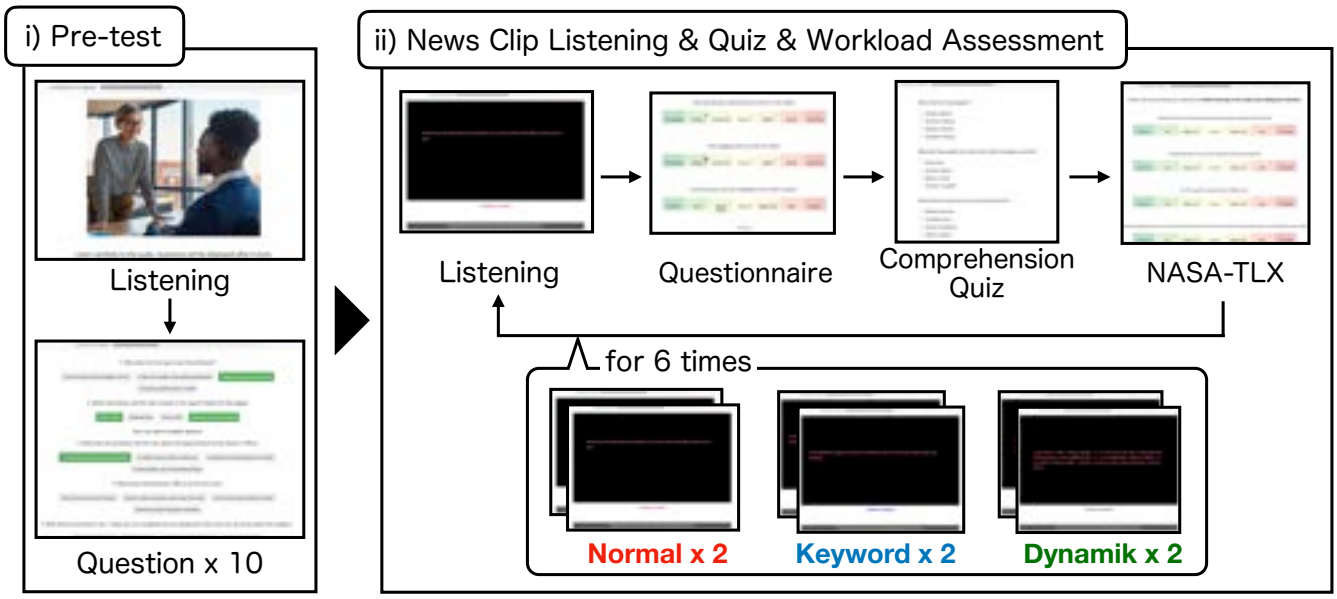


Figure 4: Main part of the experiment workflow. i) Participants listened to an audio track and completed 10 TOEFL-adapted listening comprehension questions (*Pre-test*) to assess their English listening skills. ii) The participants then listened to six CNN news clips [107] with assistive subtitles (either one from *Normal*, *Keyword*, or *Dynamik*). After each excerpt, participants completed three questionnaires to ask their self-awareness of the extent of engagement with watching the clip, their self-awareness of the extent of comprehension of the clip, and the readability of the subtitle during the clip. After that, they also completed NASA–TLX assessments and listening comprehension quizzes (*Comprehension Quiz*) on the clip [40]. This step is repeated six times with the randomized order of the video clips (two clips for every condition).

5.3.1 Additional TOEFL Questions. Here's the prompt we used to create additional TOEFL questions: "Please create an additional question for the following TOEFL statement. Make sure that the questions do not cover the same subject matter as the original question. The original question is as follows: {*Original Questions*}"

5.3.2 Quiz Validity. To assess listening comprehension of the news clip, we generated quizzes (*Comprehension Quiz*) using AI models (Claude 3.5 Sonnet [3], Gemini 1.5 Pro [34], and ChatGPT 4o [70]). The prompt we used to generate additional *Comprehension Quiz* was "Attached are 6 independent CNN news transcriptions. For each of these news items, create three quizzes to test your understanding of the content. The quiz has four choices for each question and only one of the choices is correct". After that, we discussed the validity of the quiz questions and the correct answer choices, make revisions of the contradictions, and then reduce the number of questions from nine to seven to ensure that they did not interfere with other questions. We attach the quizzes and the answer in the appendix. The information of the *Comprehension Quiz* is shown in the Appendix (See Table 3, Table 4, Table 5, Table 6, Table 7 and Table 8).

6 Result

Here, we discuss the results of each metric in the experiment.

6.1 Distribution of Conditions for Each News Clip

Figure 5 shows the condition distribution for each news clip. Since we assigned each participant to watch every clip, each video was watched exactly 84 times and there was no significant bias on the distribution by conditions.

6.2 Demographics

Figure 6 shows the demographic data of the participants in this experiment. We recruited participants whose native languages are full of variety (Figure 6 (A), (D)). We recruited 18 English native speakers as well as English non-native speakers to measure the validity of the *Comprehension Quiz* and *Pre-test*. Most of the participants are bilingual or trilingual (Figure 6 (B)), and their mother tongues are divided into European and some local languages from Asian countries (such as Japanese, Chinese, Hindi) (Figure 6 (C)).

Table 1 shows the demographic data of the participants by groups that we investigated. We divide demographics according to the following : All participants, people whose *Pre-test* scores were seven or lower (*Pre-test* ≤ 7), people whose *Pre-test* scores were above seven (*Pre-test* > 7), English Non-Natives, and English Natives. The pretest ≤ 7 and English Non-Natives were assumed to have almost the same meaning, as well as the pretest > 7 and English natives. However, we noticed that some people who declared that they are native English speakers have lower scores on the *Pre-test* than some of the people who reported that they are non-native English

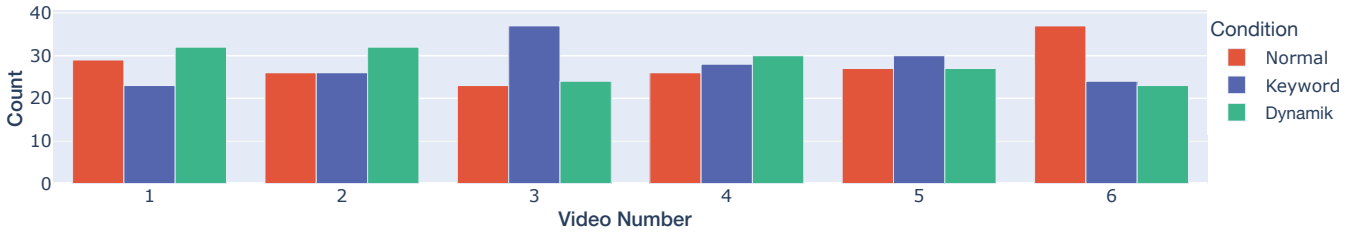


Figure 5: Condition distribution for each video news clip.

Table 1: Demographic data of all groups we have investigated.

Characteristic		All Participants	Pre-test ≤ 7	Pre-test >7	English Non-Native	English Native
Age	mean	30.1	31.5	29.3	30.1	30.1
	median	27.0	27.0	27.5	26.5	29.0
	range	20–67	20–67	20–53	20–67	20–57
Gender	man	51 (60.7 %)	18 (60.0 %)	33 (61.1 %)	41 (62.1 %)	10 (55.6 %)
	non-binary	3 (3.6 %)	2 (6.7 %)	1 (1.9 %)	3 (4.5 %)	-
	woman	30 (35.7 %)	10 (33.3 %)	20 (37.0 %)	22 (33.3 %)	8 (44.4 %)
Pre-test Score	mean	7.42	4.77	8.89	7.05	8.78
	range	2–10	2–7	8–10	2–10	7–10
Nationality	1st	Mexico: 10 (11.9 %)	Japan: 5 (16.7 %)	South Africa: 8 (14.8 %)	Mexico: 10 (15.2 %)	South Africa: 9 (50.0 %)
	2nd	Portugal: 9 (10.7 %)	Mexico: 3 (10.0 %)	Mexico: 7 (13.0 %)	Portugal: 8 (12.1 %)	United Kingdom: 5 (27.8 %)
	3rd	South Africa: 9 (10.7 %)	Portugal: 3 (10.0 %)	Portugal: 6 (11.1 %)	Poland: 8 (12.1 %)	Kenya: 2 (11.1 %)
Native Language	1st	English: 18 (21.4 %)	Portuguese: 5 (16.7 %)	English: 17 (31.5 %)	Spanish: 12 (18.2 %)	English: 18 (100.0 %)
	2nd	Spanish: 12 (14.3 %)	Japanese: 5 (16.7 %)	Spanish: 9 (16.7 %)	Portuguese: 10 (15.2 %)	-
	3rd	Portuguese: 10 (11.9 %)	Spanish: 3 (10.0 %)	Chinese: 6 (11.1 %)	Polish: 8 (12.1 %)	-

speakers. Therefore, we investigated both to evaluate our system by participants' capabilities and cultural factors.

6.3 All Participants

6.3.1 Comprehension, Engagement, and Readability. *Comprehension*, *Engagement*, and *Readability* in Figure 7 show the results of all participants' self-awareness of their comprehension of the video clips, engagement with the clips, and the readability of the subtitles (We asked "How well did you understand the content of the video?", "How engaged were you with the video?", and "How would you rate the readability of the video content?").

6.3.2 Workload. *Mental Demand*, *Physical Demand*, *Temporal Demand*, *Performance*, *Effort*, and *Frustration* in Figure 7 show the participants' self-awareness of the workload during their listening

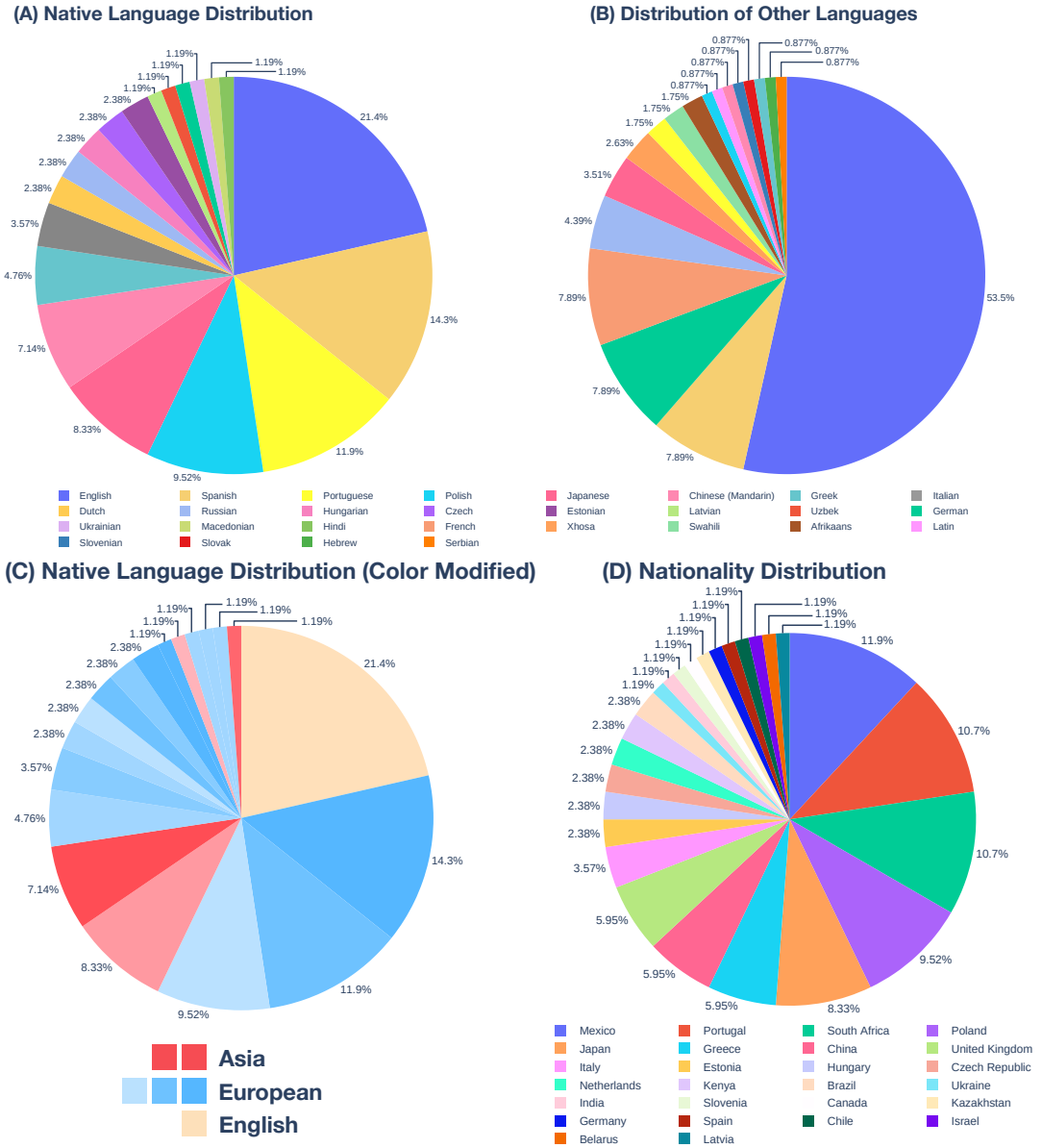


Figure 6: Demographic distribution of the survey. Note that the color matches the language in graphs A) and B) but not in graphs C) and D). A) Distribution of native languages B) Distribution of other languages that the participants can speak. C) Native language distribution, the same as A), but we modified its color according to the language families. D) Distribution of nationality.

to the news and reading the subtitles. We attached each statistical test's result in the Appendix (see Table 9, Table 10, Table 11, Table 12, and Table 13).

6.3.3 Comprehension Quizzes. '**Num Correct**' in Figure 7 shows the number of correct answers to *Comprehension Quiz* by each condition.

6.4 Other Groups

Figure 8 and Figure 9 are the result of the group whose *Pre-test* scores are seven or lower and the group whose *Pre-test* scores are greater than seven. The tendency of each items are the same as those of all participants in the *Pre-test* ≤ 7, but not in the *Pre-test* > 7. The same can be said for non-native English-speaking people and natives (see Figure 10 and Figure 11 in the appendix).

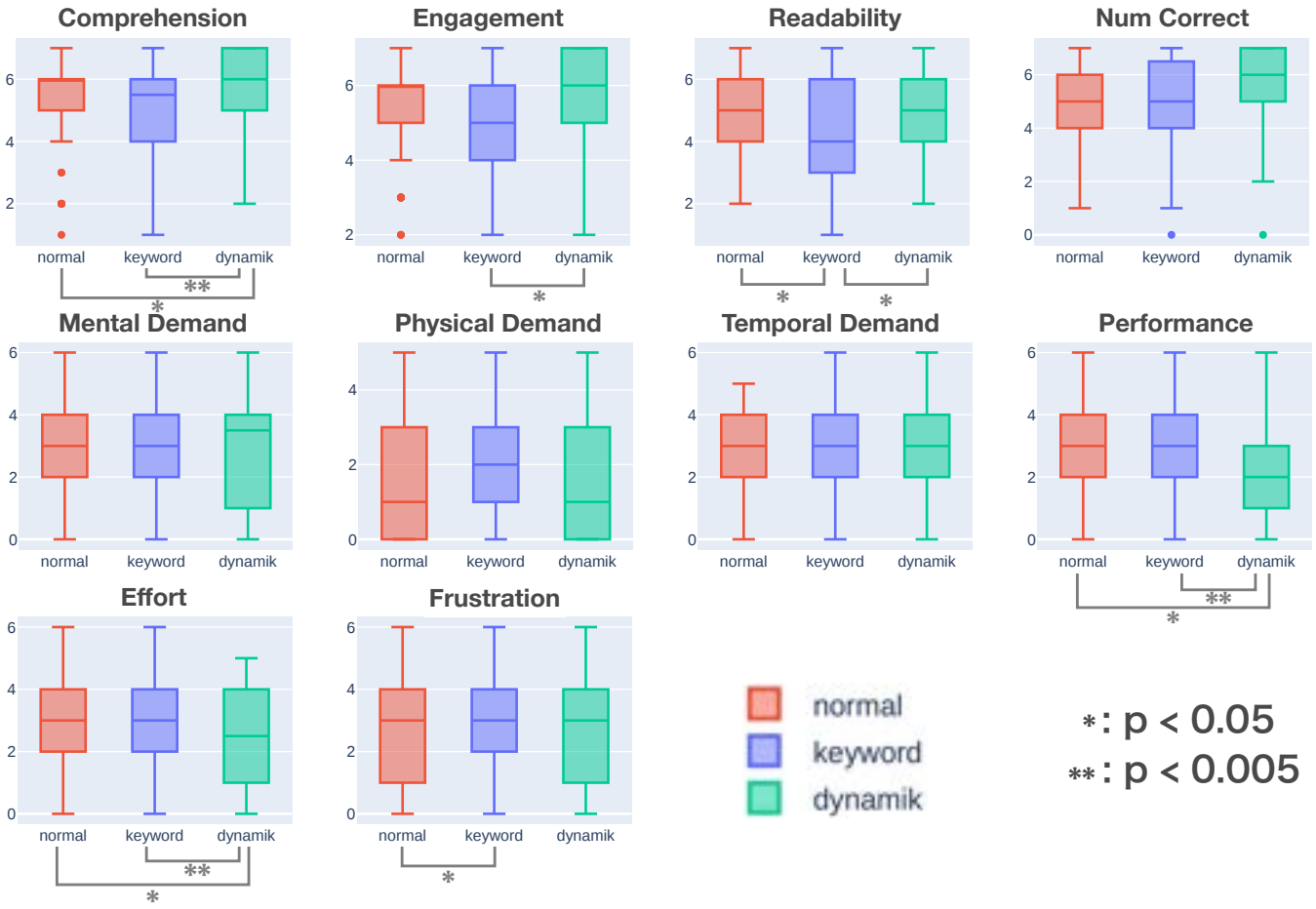


Figure 7: Workload and each item on the questionnaire of all participants.

7 Discussion

7.1 Validity of AI-Generated Quizzes

We used generative AI to create quizzes to assess listening comprehension. Manual corrections were necessary, addressing issues such as unbalanced answer choice, numerical calculation errors, logical inconsistencies, and verbatim extractions from the source text. This approach of using AI for quiz generation could be applied to other areas, such as foreign language learning. Future improvements should focus on these specific areas of weakness.

7.2 Comprehension Quiz Performance

The results of *Comprehension Quiz* Performance suggest that there were no significant differences in the average number of correct answers (*Num Correct*) across subtitle conditions, except for the case between *Keyword* and *Dynamik* conditions in the group of people whose *Pre-test* scores are seven or less. This might be because the quizzes test memory recall more than comprehension, suggesting that while assistive subtitles may aid reading, they may not significantly impact information retention. In fact, some participants indicated this; e.g.,"this didn't really test listening comprehension as much as memory".

There were also people who did not notice how the different font sizes related to the importance of the word, although we instructed them in the experiment. Although there is a limitation of data clensing on participants who were not engaged in the experiment through crowdsourcing, if we could narrow down participants to whom they fully understood the instruction and committed without skipping it, there is the possibility of a significant difference of the *Comprehension Quiz* results by the conditions.

7.3 Self-Awareness of the Extent of Comprehension, Readability, and Engagement: Are Non-native Speakers Same as English Dyslexia?

Regarding self-awareness of comprehension, readability of the texts, and engagement in reading the text, we discuss by group on *Pre-test* scores.

In the group of participants whose *Pre-test* scores are above seven, there were no significant differences.

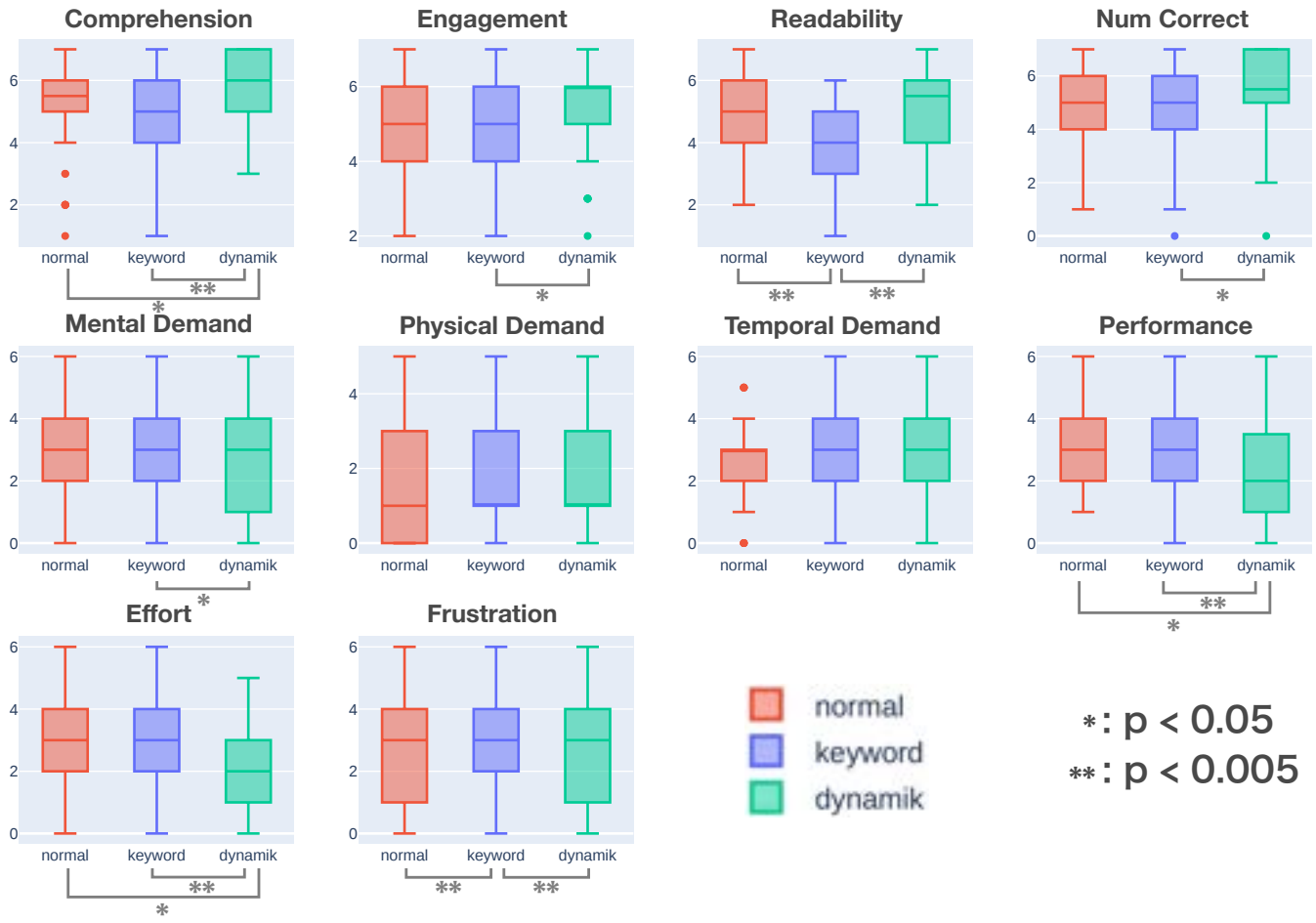


Figure 8: Workload and each item on the questionnaire of participants whose *Pre-test* scores are seven or lower.

However, in the group of participants whose *Pre-test* scores are seven or lower, while the *Comprehension* scores showed significant differences, the *Readability* scores and *Engagement* scores did not. This means that although readability and concentration level while reading were not significantly affected by text conditions, many participants felt that the *Dynamik* subtitles helped improve their comprehension.

This aligns with previous research on dyslexia, suggesting that non-native language readers might have similar characteristics to those with dyslexia when processing text [86], experiencing similar challenges. If this parallel holds true, methods and services developed for dyslexic individuals could potentially benefit non-native language learners.

As an additional implication of the relativity of these two features, previous studies have shown that other features, such as gaze, have differentiated native / non-native speakers or people with / without dyslexia [31, 81]. In future work, we will investigate the potential correlation of the characteristics of language fluency and dyslexia.

7.4 NASA–TLX Results

Here, we also discuss by group on *Pre-test* scores.

As in the same case as *Comprehension*, *Readability* and *Engagement*, in the group of participants whose *Pre-test* scores are above seven, there were no significant differences.

However, in the group of participants whose *Pre-test* scores are seven or lower, significant differences ($p < 0.05$) were observed in *Performance* and *Effort*.

This suggests that *Dynamik* facilitates people's sense of performance of reading the sentences, as is the sense of comprehension, and it alleviates people's cognitive workload, in the case when people have relatively low English skills.

7.5 Applicability to Other Languages

Although our method could be adapted to other languages, considerations for language-specific characteristics are necessary. For example, Japanese, with its three writing systems, might not benefit as much from size-based emphasis due to the inherent visual cues provided by the kanji characters, which means kanji has more meaning density compared to the alphabet [66].

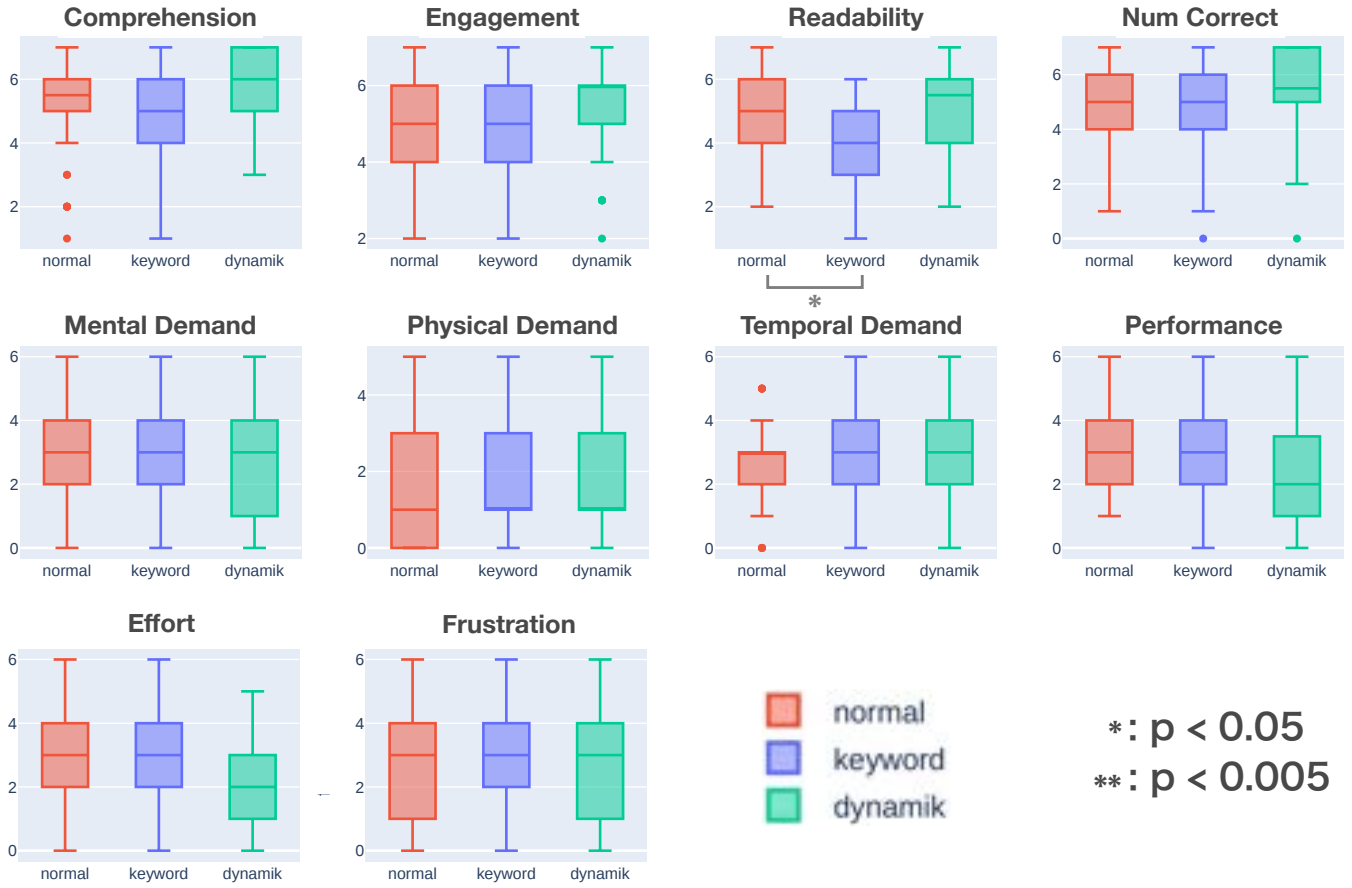


Figure 9: Workload and each item on the questionnaire of all participants with *Pre-test* scores above seven.

7.6 Font Size Variation

The size variation in *Dynamik* subtitles might have increased visual stimulation, potentially affecting NASA–TLX results. In this research, we examined pairs of 12 pt and 18 pt because they offer optimal readability for a wide audience, including those with dyslexia and the elderly [9, 10, 83–85]. However, future studies should explore the ideal ratio of sizes for better results.

7.7 Potential for Reducing Subtitle Display Area

By categorizing words into content and function words, we reduce the size of approximately 40 % of the words. With function words displayed at 2/3 the size of content words, our method resulted in subtitles occupying about 80 % of the original length ($1.0 \times 60\,\% + 0.67 \times 40\,\% \approx 80\,\%$). This reduction in occupied area could be beneficial, especially for devices with limited display space.

7.8 Alternative Display Methods

Based on participant feedback, alternative methods such as using transparency, bold text, color changes, or multilevel size adjustments could be explored in future studies. Our focus on font size was driven by the practical benefit of reducing subtitle area, but other visual cues might prove effective as well. Considering that the tendency of non-native speakers' senses of *Readability*, *Engagement* and *Comprehension* is similar to that of people with dyslexia, other display methods that are useful for people with dyslexia will also work, such as boldness [86].

In addition, according to the free-form questionnaire, some people were struggling to use *Dynamik*, saying that they did not notice what the difference in font size meant at first, or it stressed them out.

Also, in this research, we did not explore patterns besides alternative display methods; font size ratio, combination of colors and fonts, and so on. More exploration is needed to explore this novice display method.

7.9 Better Keyword Extraction Method

Although we used spaCy for our current implementation, some participants noted discrepancies between expected important words and displayed keywords, some participants saying "important keywords differ depending on each person". This suggests limitations in classification based solely on parts of speech.

In this study, we broadly categorized words into function words and content words; however, a more detailed classification could provide deeper insights. For instance, proper nouns may carry

greater importance than common nouns. Additionally, future improvements could involve light-weight machine learning models, potentially training them to predict the importance of each word (e.g., tf-idf values) for upcoming words. One idea is that reinforcement learning could be employed to develop AI models that mimic human gaze patterns and cognitive behaviors, serving as a tool for more comprehensive evaluation and training models.

7.10 System Latency

Morphological analysis introduces a delay of 0.2–0.3 seconds for space-delimited languages like English and up to 0.5 seconds for languages without spacing. The Markov process used requires processing a chunk of text from the beginning, which adds to the lag. In our real-time system, this resulted in a delay of about 0.5 seconds in displaying speech recognition results. Although this delay was consistent in the recorded videos used for crowd-sourcing, some participants still noted the lag.

8 Conclusion

Our study investigated the effectiveness of the *Dynamik* subtitle method, which emphasizes content words and de-emphasizes function words through font size. The results showed that it significantly improves the user's self-awareness of comprehension and reduces some cognitive load (*Effort* and *Performance*), especially among non-native English speakers with effect sizes of huge (> 0.8). As global content consumption continues to grow, we believe that research on listening aid methods, such as this research, is crucial to improve cross-cultural communication.

Acknowledgments

This work was supported by JST BOOST Grant JPMJBS2418, JST Moonshot R&D Grant JPMJMS2012, JST CREST Grant JPMJCR17A3, and the commissioned research by NICT Japan Grant JPJ012368C02901.

References

[1] Antar Solhy Abdellah. 2008. Can intralingual subtitling enhance English major's listening comprehension of literary texts? *Journal of Qena Faculty of Education* 11, 11 (2008), 116–162.

[2] Adobe Firefly 2024. Adobe Firefly - Free Generative AI for creatives. https://www.adobe.com/products/firefly.html. (Accessed on 09/12/2024).

[3] Anthropic [n. d.]. Claude. https://www.anthropic.com/news/claude-3-5-sonnet. (Accessed on 09/12/2024).

[4] azure 2024. Azure AI Speech. (Accessed on 09/12/2024).

[5] azure_sdk 2024. Azure SDK for .NET. (Accessed on 09/12/2024).

[6] Keith Bain, Sara H. Basson, and Mike Wald. 2002. Speech recognition in university classrooms: liberated learning project. In *Proceedings of the Fifth International ACM Conference on Assistive Technologies* (Edinburgh, Scotland) *(Assets '02)*. Association for Computing Machinery, New York, NY, USA, 192–196. https://doi.org/10.1145/638249.638284

[7] Stephanie Berger, Oliver Niebuhr, and Kerstin Fischer. 2018. Eliciting extra prominence in read-speech tasks: The effects of different text-highlighting methods on acoustic cues to perceived prominence. In *Proceedings of the 9th International Conference on Speech Prosody*. 75–79. https://doi.org/10.21437/SpeechProsody.2018-15

[8] Larwan Berke, Christopher Caulfield, and Matt Huenerfauth. 2017. Deaf and Hard-of-Hearing Perspectives on Imperfect Automatic Speech Recognition for Captioning One-on-One Meetings. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 155–164. https://doi.org/10.1145/3132525.3132541

[9] Michael Bernard, Chia Hui Liao, and Melissa Mills. 2001. The effects of font type and size on the legibility and reading time of online text by older adults. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems* (Seattle, Washington) *(CHI EA '01)*. Association for Computing Machinery, New York, NY, USA, 175–176. https://doi.org/10.1145/634067.634173

[10] Sanjiv K Bhatia, Ashok Samal, Nithin Rajan, and Marc T Kiviniemi. 2011. Effect of font size, italics, and colour count on web usability. *International journal of computational vision and robotics* 2, 2 (2011), 156–179.

[11] Florian Boudin. 2018. Unsupervised Keyphrase Extraction with Multipartite Graphs. arXiv:1803.08721 [cs.IR] https://arxiv.org/abs/1803.08721

[12] Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Ruslan Mitkov and Jong C. Park (Eds.). Asian Federation of Natural Language Processing, Nagoya, Japan, 543–551. https://aclanthology.org/I13-1062

[13] British Council, IDP IELTS, and Cambridge University Press & Assessment 2024. IELTS. https://ielts.org/. (Accessed on 09/12/2024).

[14] CYRIL BURT, W. F. COOPER, and J. L. MARTIN. 1955. A PSYCHOLOGICAL STUDY OF TYPOGRAPHY. *British Journal of Statistical Psychology* 8, 1 (1955), 29–56. https://doi.org/10.1111/j.2044-8317.1955.tb00160.x arXiv:https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8317.1955.tb00160.x

[15] Annamaria Caimi. 2006. Audiovisual Translation and Language Learning: The Promotion of Intralingual Subtitles. *The Journal of Specialised Translation Issue* 6, 85–98. https://api.semanticscholar.org/CorpusID:61372626

[16] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE: Keyword extraction from single documents using multiple local features. *Information Sciences* 509 (2020), 257–289. https://doi.org/10.1016/j.ins.2019.09.013

[17] Rudolf Carnap. 1937. *The Logical Syntax of Language*. K. Paul, Trench, Trubner & co., London.

[18] csharp 2024. C# Language Design. (Accessed on 09/12/2024).

[19] Richard L. Daft and Robert H. Lengel. 1986. Organizational Information Requirements, Media Richness and Structural Design. *Manage. Sci.* 32, 5 (may 1986), 554–571. https://doi.org/10.1287/mnsc.32.5.554

[20] Caluã de Lacerda Pataca, Saad Hassan, Nathan Tinker, Roshan Lalintha Peiris, and Matt Huenerfauth. 2024. Caption Royale: Exploring the Design Space of Affective Captions from the Perspective of Deaf and Hard-of-Hearing Individuals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 899, 17 pages. https://doi.org/10.1145/3613904.3642258

[21] Joshua R. de Leeuw, Rebecca A. Gilbert, and Björn Luchterhandt. 2023. jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments. *Journal of Open Source Software* 8, 85 (2023), 5351. https://doi.org/10.21105/joss.05351

[22] dotnet 2024. .NET 8.0. (Accessed on 09/12/2024).

[23] Educational Testing Service 2024. TOEFL. https://www.ets.org/toefl.html. (Accessed on 09/12/2024).

[24] Educational Testing Service 2024. TOEIC. https://www.ets.org/toeic.html. (Accessed on 09/12/2024).

[25] Samhaa R. El-Beltagy and Ahmed Rafea. 2010. KP-Miner: Participation in SemEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, Katrin Erk and Carlo Strapparava (Eds.). Association for Computational Linguistics, Uppsala, Sweden, 190–193. https://aclanthology.org/S10-1041

[26] Marco Porta Elisa Perego, Fabio Del Missier and Mauro Mosconi. 2010. The Cognitive Effectiveness of Subtitle Processing. *Media Psychology* 13, 3 (2010), 243–272. https://doi.org/10.1080/15213269.2010.502873

[27] en_core_web_sm 2024. en_core_web_sm. (Accessed on 09/12/2024).

[28] Corina Florescu and Cornelia Caragea. 2017. PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 1105–1115. https://doi.org/10.18653/v1/P17-1102

[29] Jodi Forlizzi, Johnny Lee, and Scott Hudson. 2003. The kinedit system: affective messages using dynamic texts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 377–384. https://doi.org/10.1145/642611.642677

[30] Charles Carpenter Fries. 1952. *The Structure of English*. Harcourt Brace & World, Inc.

[31] Katsuya Fujii and Jun Rekimoto. 2019. SubMe: An Interactive Subtitle System with English Skill Estimation Using Eye Tracking. In *Proceedings of the 10th Augmented Human International Conference 2019* (Reims, France) *(AH2019)*. Association for Computing Machinery, New York, NY, USA, Article 23, 9 pages. https://doi.org/10.1145/3311823.3311865

[32] Olivia Gerber-Morón, Olga Soler-Vilageliu, Judit Castellà, et al. 2020. Effects of screen size on subtitle layout preferences and comprehension across devices. (2020).

[33] Gitgub. [n. d.]. Webpack. https://github.com/webpack/webpack. (Accessed on 09/12/2024).

[34] Google [n. d.]. Gemini. https://gemini.google.com/. (Accessed on 09/12/2024).

[35] Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. https://doi.org/10.5281/zenodo.4461265

[36] Shashi Kant Gupta. 2024. Brain2AI/jsPsychSheet: jsPsychSheet: A simple JavaScript library that uses jsPsych and Google Sheet for running behavioral experiments online. (Accessed on 09/12/2024).

[37] Ting-Chia Hsu Gwo-Jen Hwang and Yi-Hsuan Hsieh. 2019. Impacts of Different Smartphone Caption/Subtitle Mechanisms on English Listening Performance and Perceptions of Students with Different Learning Styles. *International Journal of Human–Computer Interaction* 35, 4-5 (2019), 333–344. https://doi.org/10.1080/10447318.2018.1543091

[38] Morris Halle and K. P. Mohanan. 1985. Segmental Phonology of Modern English. *Linguistic Inquiry* 16, 1 (1985), 57–116. http://www.jstor.org/stable/4178420

[39] Michael Halliday. 1985. *Spoken and Written Language.* Deakin University Press.

[40] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[41] Ari Hautasaari and Naomi Yamashita. 2014. Catching up in audio conferences: highlighting keywords in ASR transcripts for non-native speakers. In *Proceedings of the 5th ACM International Conference on Collaboration across Boundaries: Culture, Distance & Technology* (Kyoto, Japan) *(CABS '14)*. Association for Computing Machinery, New York, NY, USA, 107–110. https://doi.org/10.1145/2631488.2634064

[42] Ari Hautasaari and Naomi Yamashita. 2014. Do automated transcripts help non-native speakers catch up on missed conversation in audio conferences?. In *Proceedings of the 5th ACM International Conference on Collaboration across Boundaries: Culture, Distance & Technology* (Kyoto, Japan) *(CABS '14)*. Association for Computing Machinery, New York, NY, USA, 65–72. https://doi.org/10.1145/2631488.2631495

[43] E Tory Higgins. 1996. Activation: Accessibility, and salience. *Social psychology: Handbook of basic principles* (1996), 133–168.

[44] John Higgins. 1983. Computer assisted language learning. *Language Teaching* 16, 2 (1983), 102–114. https://doi.org/10.1017/S0261444800009988

[45] Richang Hong, Meng Wang, Mengdi Xu, Shuicheng Yan, and Tat-Seng Chua. 2010. Dynamic captioning: video accessibility enhancement for hearing impairment. In *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. Association for Computing Machinery, New York, NY, USA, 421–430. https://doi.org/10.1145/1873951.1874013

[46] InVision. 2024. Principles of design. https://www.invisionapp.com/defined/principles-of-design.

[47] Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28 (1972), 11–21. https://api.semanticscholar.org/CorpusID:2996187

[48] Sushant Kafle, Peter Yeung, and Matt Huenerfauth. 2019. Evaluating the Benefit of Highlighting Key Words in Captions for People who are Deaf or Hard of Hearing. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 43–55. https://doi.org/10.1145/3308561.3353781

[49] Slava Kalyuga. 2011. Cognitive Load Theory: Implications for Affective Computing., In Proceedings of the 24th International Florida Artificial Intelligence Research Society. *Proceedings of the 24th International Florida Artificial Intelligence Research Society, FLAIRS - 24*, 105–110.

[50] M. H. Kelly. 1992. Using sound to solve syntactic problems: the role of phonology in grammatical category assignments. *Psychological Review* 99 (1992), 349–364. Issue 2. https://doi.org/10.1037/0033-295x.99.2.349

[51] JooYeong Kim, SooYeon Ahn, and Jin-Hyuk Hong. 2023. Visible Nuances: A Caption System to Visualize Paralinguistic Speech Cues for Deaf and Hard-of-Hearing Individuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 54, 15 pages. https://doi.org/10.1145/3544548.3581130

[52] Muriel R. Schulz Klammer, Thomas and Angela Della Volpe. 2009. *Analyzing English Grammar (6th Edition).* Longman.

[53] Raja S. Kushalnagar, Gary W. Behm, Aaron W. Kelstone, and Shareef Ali. 2015. Tracked Speech-To-Text Display: Enhancing Accessibility and Readability of Real-Time Speech-To-Text. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (Lisbon, Portugal) *(ASSETS '15)*. Association for Computing Machinery, New York, NY, USA, 223–230. https://doi.org/10.1145/2700648.2809843

[54] Raja S. Kushalnagar, Walter S. Lasecki, and Jeffrey P. Bigham. 2014. Accessibility Evaluation of Classroom Captions. *ACM Trans. Access. Comput.* 5, 3, Article 7 (jan 2014), 24 pages. https://doi.org/10.1145/2543578

[55] DANIEL G. LEE, DEBORAH I. FELS, and JOHN PATRICK UDO. 2007. Emotive captioning. *Comput. Entertain.* 5, 2, Article 11 (apr 2007), 15 pages. https://doi.org/10.1145/1279540.1279551

[56] Joshua Leeuw. 2023. DataPipe: Born-open data collection for online experiments. *Behavior Research Methods* 56 (06 2023). https://doi.org/10.3758/s13428-023-02161-x

[57] Michael Levy. 1997. *Computer-Assisted Language Learning: Context and Conceptualisation.* Oxford University Press.

[58] Clive Lewis and Peter Walker. 1989. Typographic influences on reading. *British Journal of Psychology* 80, 2 (1989), 241–257. https://doi.org/10.1111/j.2044-8295.1989.tb02317.x arXiv:https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8295.1989.tb02317.x

[59] Yuanhua Lv and ChengXiang Zhai. 2011. Lower-bounding term frequency normalization. In *International Conference on Information and Knowledge Management*. 7–16. https://api.semanticscholar.org/CorpusID:14029221

[60] Kelly Mack, Emma McDonnell, Dhruv Jain, Lucy Lu Wang, Jon E. Froehlich, and Leah Findlater. 2021. What Do We Mean by "Accessibility Research"? A Literature Survey of Accessibility Papers in CHI and ASSETS from 1994 to 2019. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 371, 18 pages. https://doi.org/10.1145/3411764.3445412

[61] Richard E. Mayer, Julie Heiser, and Steven Lonn. 2001. Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology* 93 (2001), 187–198. https://api.semanticscholar.org/CorpusID:28827894

[62] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Dekang Lin and Dekai Wu (Eds.). Association for Computational Linguistics, Barcelona, Spain, 404–411. https://aclanthology.org/W04-3252

[63] Mohammad Reza Mirzaei, Seyed Ghorshi, and Mohammad Mortazavi. 2012. Combining Augmented Reality and Speech Technologies to Help Deaf and Hard of Hearing People. In *14th Symposium on Virtual and Augmented Reality*. 174–181. https://doi.org/10.1109/SVR.2012.10

[64] Louisa Cook Moats and G. Reid Lyon. 1993. Learning Disabilities in the United States: Advocacy, Science, and the Future of the Field. *Journal of Learning Disabilities* 26, 5 (1993), 282–294. https://doi.org/10.1177/002221949302600501 PMID: 8492047.

[65] Hidetaka Nambo, Shuichi Seto, Hiroshi Arai, Kimikazu Sugimori, Yuko Shimomura, and Hiroyuki Kawabe. 2012. Visualization of Non-verbal Expressions in Voice for Hearing Impaired. In *COMPUTERS HELPING PEOPLE WITH SPECIAL NEEDS, PT I*. 492–499. https://doi.org/10.1007/978-3-642-31522-0_74

[66] Naoyuki Osaka Natsumi Kajii, Tatjana A. Nazir. 2001. Eye movement control in reading unspaced text: the case of the Japanese script. *Vision Research* 41, 19 (2001), 2503–2510.

[67] Node.js developers [n. d.]. Node.js. https://nodejs.org/en. (Accessed on 09/12/2024).

[68] npm developers [n. d.]. npm. https://www.npmjs.com/. (Accessed on 09/12/2024).

[69] Kotaro Oomori, Akihisa Shitara, Tatsuya Minagawa, Sayan Sarcar, and Yoichi Ochiai. 2020. A Preliminary Study on Understanding Voice-only Online Meetings Using Emoji-based Captioning for Deaf or Hard of Hearing Users. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 54, 4 pages. https://doi.org/10.1145/3373625.3418032

[70] OpenAI [n. d.]. ChatGPT. https://openai.com/chatgpt/. (Accessed on 09/12/2024).

[71] Jason Orlosky, Kiyoshi Kiyokawa, and Haruo Takemura. 2014. Managing mobile text in head mounted displays: studies on visual preference and text placement. *SIGMOBILE Mob. Comput. Commun. Rev.* 18, 2 (jun 2014), 20–31. https://doi.org/10.1145/2636242.2636246

[72] osf 2024. Open Science Framework. (Accessed on 09/12/2024).

[73] Fred Paas, Alexander Renkl, and John Sweller. 2004. Cognitive Load Theory: Instructional Implications of the Interaction between Information Structures and Cognitive Architecture. *Instructional Science* 32 (2004), 1–8. https://api.semanticscholar.org/CorpusID:6978902

[74] Mei-Hua Pan, Naomi Yamashita, and Hao-Chuan Wang. 2017. Task Rebalancing: Improving Multilingual Communication with Native Speakers-Generated Highlights on Automated Transcripts. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 310–321. https://doi.org/10.1145/2998181.2998304

[75] Hector R. Ponce and Richard E. Mayer. 2014. An eye movement analysis of highlighting and graphic organizer study aids for learning from expository text. *Computers in Human Behavior* 41 (2014), 21–32. https://doi.org/10.1016/j.chb.2014.09.010

[76] Prolific [n. d.]. Prolific. https://www.prolific.com. (Accessed on 09/12/2024).

[77] python 2024. Python. (Accessed on 09/12/2024).

[78] Luz Rello and Ricardo Baeza-Yates. 2013. Good fonts for dyslexia. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and*

Accessibility (Bellevue, Washington) *(ASSETS '13)*. Association for Computing Machinery, New York, NY, USA, Article 14, 8 pages. https://doi.org/10.1145/2513383.2513447

[79] Luz Rello and Ricardo Baeza-Yates. 2016. The Effect of Font Type on Screen Readability by People with Dyslexia. *ACM Trans. Access. Comput.* 8, 4, Article 15 (may 2016), 33 pages. https://doi.org/10.1145/2897736

[80] Luz Rello and Ricardo A. Baeza-Yates. 2015. How to present more readable text for people with dyslexia. *Universal Access in the Information Society* 16 (2015), 29–49. https://api.semanticscholar.org/CorpusID:35288203

[81] Luz Rello and Miguel Ballesteros. 2015. Detecting readers with dyslexia using machine learning with eye tracking measures. In *Proceedings of the 12th International Web for All Conference* (Florence, Italy) *(W4A '15)*. Association for Computing Machinery, New York, NY, USA, Article 16, 8 pages. https://doi.org/10.1145/2745555.2746644

[82] Luz Rello and Jeffrey P. Bigham. 2017. Good Background Colors for Readers: A Study of People with and without Dyslexia. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 72–80. https://doi.org/10.1145/3132525.3132546

[83] Luz Rello, Gaurang Kanvinde, and Ricardo Baeza-Yates. 2012. Layout guidelines for web text and a web service to improve accessibility for dyslexics. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility* (Lyon, France) *(W4A '12)*. Association for Computing Machinery, New York, NY, USA, Article 36, 9 pages. https://doi.org/10.1145/2207016.2207048

[84] Luz Rello, Martin Pielot, and Mari-Carmen Marcos. 2016. Make It Big! The Effect of Font Size and Line Spacing on Online Readability. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3637–3648. https://doi.org/10.1145/2858036.2858204

[85] Luz Rello, Martin Pielot, Mari-Carmen Marcos, and Roberto Carlini. 2013. Size matters (spacing not): 18 points for a dyslexic-friendly Wikipedia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility* (Rio de Janeiro, Brazil) *(W4A '13)*. Association for Computing Machinery, New York, NY, USA, Article 17, 4 pages. https://doi.org/10.1145/2461121.2461125

[86] Luz Rello, Horacio Saggion, and Ricardo Baeza-Yates. 2014. Keyword Highlighting Improves Comprehension for People with Dyslexia. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, Sandra Williams, Advaith Siddharthan, and Ani Nenkova (Eds.). Association for Computational Linguistics, Gothenburg, Sweden, 30–37. https://doi.org/10.3115/v1/W14-1204

[87] Ronald E. Rice. 2006. Media Appropriateness: Using Social Presence Theory to Compare Traditional and New Organizational Media. *Human Communication Research* 19, 4 (03 2006), 451–484. https://doi.org/10.1111/j.1468-2958.1993.tb00309.x arXiv:https://academic.oup.com/hcr/article-pdf/19/4/451/22342350/jhumcom0451.pdf

[88] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (apr 2009), 333–389. https://doi.org/10.1561/1500000019

[89] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management* (Washington, D.C., USA) *(CIKM '04)*. Association for Computing Machinery, New York, NY, USA, 42–49. https://doi.org/10.1145/1031171.1031181

[90] Tim Schopf, Simon Klimek, and Florian Matthes. 2022. PatternRank: Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase Extraction. In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0011546600003335

[91] Alina Secara. 2011. R U ready 4 new subtitles? Investigating the potential of social translation practices and creative spellings. *Community Translation 2.0* 10 (10 2011), 1–19. https://doi.org/10.52034/lanstts.v10i.282

[92] Educational Testing Service. 2020. *The Official Guide to the TOEFL iBT Test (5th Edition)*. McGraw-Hill.

[93] Prafull Sharma and Yingbo Li. 2019. Self-Supervised Contextual Keyword and Keyphrase Retrieval with Self-Labelling. https://api.semanticscholar.org/CorpusID:199535532

[94] Sally E Shaywitz, Michael D Escobar, Bennett A Shaywitz, Jack M Fletcher, and Robert Makuch. 1992. Evidence that dyslexia may represent the lower tail of a normal distribution of reading ability. *New England Journal of Medicine* 326, 3 (1992), 145–150.

[95] spacy 2024. spaCy. (Accessed on 09/12/2024).

[96] Lucas Sterckx, Thomas Demeester, Johannes Deleu, and Chris Develder. 2015. Topical Word Importance for Fast Keyphrase Extraction. *Proceedings of the 24th International Conference on World Wide Web* (2015). https://api.semanticscholar.org/CorpusID:2431019

[97] Hendrik Strobelt, Daniela Oelke, Bum Chul Kwon, Tobias Schreck, and Hanspeter Pfister. 2016. Guidelines for Effective Usage of Text Highlighting Techniques. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (jan 2016), 489–498. https://doi.org/10.1109/TVCG.2015.2467759

[98] Peter Turney. 2002. Learning to Extract Keyphrases from Text. *CoRR* cs.LG/0212013 (12 2002), 1–45.

[99] unity 2024. Unity. (Accessed on 09/12/2024).

[100] universalcolor 2024. Model Color Palette for Color Universal Design Guide Book. (Accessed on 09/12/2024).

[101] S. Nagasawa V. Ferdiansyah. 2013. Effect of captioning lecture videos for learning in foreign language. In *Proceedings of SLP Meeting of Information Processing Society of Japan*.

[102] Larry Vandergrift. 2007. Recent Developments in Decond and Foreign Language Listening Comprehension Research. *Language Teaching* 40, 3 (2007), 191–210. https://doi.org/10.1017/S0261444807004338

[103] Keith Vertanen and Per Ola Kristensson. 2008. On the benefits of confidence visualization in speech recognition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1497–1500. https://doi.org/10.1145/1357054.1357288

[104] Vue.js developers. [n. d.]. Vue.js. https://vuejs.org/. (Accessed on 09/12/2024).

[105] Xiaojun Wan and Jianguo Xiao. 2008. CollabRank: Towards a Collaborative Approach to Single-Document Keyphrase Extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Donia Scott and Hans Uszkoreit (Eds.). Coling 2008 Organizing Committee, Manchester, UK, 969–976. https://aclanthology.org/C08-1122

[106] Fangzhou Wang, Hidehisa Nagano, Kunio Kashino, and Takeo Igarashi. 2017. Visualizing Video Sounds With Sound Word Animation to Enrich User Experience. *IEEE Transactions on Multimedia* 19, 2 (2017), 418–429. https://doi.org/10.1109/TMM.2016.2613641

[107] Warner Bros. Discovery 2024. Cable News Network. https://edition.cnn.com/. (Accessed on 09/12/2024).

[108] wcag 2024. Web Content Accessibility Guidelines (WCAG) 2.1. (Accessed on 09/12/2024).

[109] Thomas Wehr and Werner Wippich. 2004. Typography and color: Effects of salience and fluency on conscious recollective experience. *Psychological research* 69 (2004), 138–146.

[110] Guande Wu, Jing Qian, Sonia Castelo Quispe, Shaoyu Chen, João Rulff, and Claudio Silva. 2024. ARTiST: Automated Text Simplification for Task Guidance in Augmented Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 939, 24 pages. https://doi.org/10.1145/3613904.3642772

[111] Xiniel Zhang Yuzu Saijo and Jun Rekimoto. 2016. Flash Word : ESL Listening Support with Subtitle Skimming. In *Proceedings of the 1st Asian Workshop on User Interface*. 1–2.

[112] Hugo Zaragoza, Nick Craswell, Michael J. Taylor, Suchi Saria, and Stephen E. Robertson. 2004. Microsoft Cambridge at TREC 13: Web and Hard Tracks. In *Text Retrieval Conference*. https://api.semanticscholar.org/CorpusID:7420138

[113] zenmarugothic. 2024. Zen Maru Gothic.

[114] zeromq 2024. ZeroMQ. (Accessed on 09/12/2024).

Table 2: Information of news clips used in the experiment.

News	Total words	Content words	Function words	Lexical Density (%)	News Title	Transcription
1	84	50	34	60	Global wildlife down 60 %	1
2	70	38	32	54	Natural mushroom cloud causes alarm	2
3	90	53	37	59	Impact of air pollution on children	3
4	60	35	25	58	LinkedIn holds workplace parents day	4
5	79	50	29	63	World rankings in kids science skills	5
6	52	32	20	62	Fake news prompts gunman's raid	6

[1]A new report from the World Wildlife Fund has found that our planet has lost almost 60 % of its wildlife in less than half a century. Scientists say the rapid extinction is caused by the loss of habitat that comes from pollution, the exploitation of resources as well as climate change. The report highlights a number of species, elephants, for example, whose numbers have dropped by 5th in just ten years. As for sharks and rays, 1/3 are threatened by overfishing.

[2]It might look like a sign of nuclear warfare. How would you feel if you saw this big mushroom cloud hanging over your neighborhood? Well, this one appeared in Western Siberia, and according to Russian media, a number of panicky people called emergency services fearing a nuclear attack. I don't blame them. Turns out this formation happens when a thunderstorm. Causes clouds to be blown sideways.

[3]Now air pollution is a serious global health concern. UNICEF says around 600,000 children under the age of five die every year from pollution related illnesses and also warns that pollutants can permanently damage children's brain development. Around 2 billion children live in places where pollution levels exceed WHO guidelines. And most of the pollution comes from burning fossil fuel and vehicle emissions. But dangers also lie at home. Around 1 billion children live in homes that use wood and coal for cooking and heating.

[4]Parents across the globe are checking up on their kids at the office right now as part of Linkedin's Bring Your Parents Day. Here you can see pictures from social media showing how the visits are turning out. In a generation where more and more jobs are becoming less traditional and more flexible, LinkedIn says one in three parents cannot describe their kids job.

[5]Asia is producing some of the world's brightest students. Every four years, 10 and 14 year olds from around the world get ranked in an international math and science study. And Singapore crushes the competition in every category. For instance, among 10 year olds in science, Singapore comes in number one. That's followed by South Korea. Japan in 3rd and then Russia. Hong Kong comes in fifth. Finland is actually the only Europe in top 10.

[6]And what started out as a conspiracy theory motivated a man to bring a gun to a pizza shop. This one right there in Washington. Police say the gunman apparently believed a fake news story online, and the bogus story falsely claimed that the pizza shop was a site of a child sex ring run by Hillary Clinton and her come.

Table 3: Information of Comprehension Quiz on video number 1.

Question Number	Question Statement	Choices	Correct Answer
1	According to the WWF report, by how much has global wildlife decreased in less than half a century?	A: About 40 % B: About 60 % C: About 75 % D: About 90 %	B
2	Over how many years has the elephant population decreased by one-fifth, according to the report?	A: 5 years B: 10 years C: 15 years D: 20 years	B
3	What fraction of sharks and rays are threatened by overfishing?	A: One-fourth B: One-third C: Half D: Two-thirds	B
4	Which organization released the report on global wildlife decline?	A: United Nations B: World Bank C: World Wildlife Fund D: Greenpeace	C
5	What time frame does the report cover for the wildlife decline?	A: Less than a quarter century B: Less than half a century C: Less than a century D: More than a century	B
6	What is not the main causes of rapid extinction mentioned in the report?	A: Natural disasters B: Habitat loss and pollution C: Hunting and poaching D: Climate change	A
7	Which specific animal group is mentioned as having lost a fifth of its population in a decade?	A: Lions B: Elephants C: Tigers D: Rhinos	B

Table 4: Information of Comprehension Quiz on video number 2.

Question Number	Question Statement	Choices	Correct Answer
1	Where did this cloud appear?	A: Western Siberia B: Eastern Siberia C: Northern Siberia D: Southern Siberia	A
2	What did many people fear when they called emergency services?	A: Forest fire B: Meteor strike C: Volcanic eruption D: Nuclear attack	D
3	What was the actual cause of this cloud formation?	A: Thunderstorm B: Factory emissions C: Military exercise D: Meteor impact	A
4	How did the news describe people's reaction to seeing the cloud?	A: Curious B: Excited C: Panicky D: Indifferent	C
5	According to the news what is the shape of the cloud?	A: Mushroom B: Tornado C: Huge potato D: Thunderstorm	A
6	How did the cloud appear to be formed?	A: Rising straight up B: Blown sideways C: Spiraling D: Dissipating quickly	B
7	What did the reporter say about people's reaction?	A: It was an overreaction B: It was understandable C: It was amusing D: It was concerning	B

Table 5: Information of Comprehension Quiz on video number 3.

Question Number	Question Statement	Choices	Correct Answer
1	According to UNICEF, how many children under the age of five die every year from pollution-related illnesses?	A: About 200000 B: About 400000 C: About 600000 D: About 800000	C
2	How many children live in areas where pollution levels exceed WHO guidelines?	A: About 500 million B: About 1 billion C: About 1.5 billion D: About 2 billion	D
3	What is mentioned as a main cause of indoor air pollution?	A: Keeping pets B: Using pesticides C: Using plastic products D: Using wood and coal for cooking and heating	D
4	What long-term effect can pollutants have on children according to UNICEF?	A: Worse IQ B: Permanent brain damage C: Less physical growth D: Worse respiratory health	B
5	What are the main sources of pollution mentioned in the news?	A: Industrial waste B: Fossil fuel burning and vehicle emissions C: Agricultural runoff D: Electronic waste	B
6	How many children live in homes using wood and coal for cooking and heating?	A: About 500 million B: About 1 billion C: About 1.5 billion D: About 2 billion	B
7	How does the news describe air pollution as a health concern?	A: Subtle issue B: Serious global concern C: Localized but serious problem D: Improving situation	B

Table 6: Information of Comprehension Quiz on video number 4.

Question Number	Question Statement	Choices	Correct Answer
1	What is the name of the event held by LinkedIn?	A: Bring Your Parents Day B: Bring Your Kids to Work Day C: Family Office Day D: LinkedIn Family Event	A
2	According to LinkedIn, what fraction of parents cannot describe their children's jobs?	A: One-fourth B: One-third C: Half D: Two-thirds	B
3	What generational change is mentioned as the background for this event?	A: Decreased independence of children B: Weakening of parent-child relationships C: Deterioration of workplace environments D: Diversification and flexibility of jobs	D
4	Where are parents checking up on their kids during this event?	A: At home B: At school C: At the office D: In public spaces	C
5	How widespread is this event according to the news?	A: Local to one city B: National event C: Global event D: Limited to tech companies	C
6	How is the event being documented?	A: Through official reports B: Via social media pictures C: By news reporters D: Through LinkedIn profiles	B
7	How old are the invited people assumed to be?	A: under 20 B: 20s C: 40s to 70s D: above 70s	C

Table 7: Information of Comprehension Quiz on video number 5.

Question Number	Question Statement	Choices	Correct Answer
1	Which country ranked first in science skills for 10-year-olds?	A: Singapore B: South Korea C: Japan D: Finland	A
2	How often is this ranking conducted?	A: Every 2 years B: Every 3 years C: Every 4 years D: Every 5 years	C
3	Which is the only European country to make it into the top 5?	A: Finland B: France C: Russia D: None of them	D
4	What age groups are included in this international study?	A: 8 and 12 year olds B: 9 and 13 year olds C: 10 and 14 year olds D: 11 and 15 year olds	C
5	What subjects are included in this international study?	A: Language and science B: Math and Language C: Math and science D: Math science and language	C
6	How did the news describe Singapore's performance?	A: Good B: Above average C: Crushes the competition D: Slightly better than others	C
7	What was Russia's ranking in the science category for 10-year-olds?	A: 1st B: 2nd C: 3rd D: 4th	D

Table 8: Information of Comprehension Quiz on video number 6.

Question Number	Question Statement	Choices	Correct Answer
1	Where did the man with a gun go?	A: A pizza shop B: A government agency C: A school D: A bank	A
2	Who was falsely accused in the fake news story?	A: Donald Trump B: Barack Obama C: Joe Biden D: Hillary Clinton	D
3	What was the fake news story about?	A: Election fraud B: Child sex trafficking ring C: Money laundering D: Espionage activities	B
4	Where was the pizza shop located?	A: New York B: Washington C: Chicago D: Los Angeles	B
5	According to the news, what motivated the man to bring a gun to the pizza shop?	A: Conspiracy theory B: Personal vendetta C: Robbery attempt D: Political protest	A
6	How did the gunman come to believe the fake news story?	A: Through a newspaper B: On television C: Online D: From a friend	C
7	What did the fake news falsely claim about Hillary Clinton?	A: She owned the pizza shop B: She ran a child sex ring C: She was hiding there D: She was selling drugs there	B

Table 9: Statistical results by each condition among all participants. For the values of comprehension, engagement, and readability are 0:Not at all – 7:Perfect on Likert Scale. The value of num correct ranges from 0 to 10. For the values of mental demand, physical demand, temporal demand, performance, effort, and frustration are 0:Perfect – 7:Worst on Likert Scale.

Group	Metric	Condition 1	Condition 2	U-statistic	p-value	Cohen's d
All Participants	comprehension	normal	keyword	3286.00	0.4306	-0.16
		normal	dynamik	4334.00	**0.0077**	0.41
		keyword	dynamik	4498.00	**0.0014**	0.55
	engagement	normal	keyword	3093.00	0.1548	-0.25
		normal	dynamik	3923.00	0.1957	0.20
		keyword	dynamik	4348.00	**0.0072**	0.44
	readability	normal	keyword	2351.50	**0.0001**	-0.64
		normal	dynamik	3417.00	0.7192	-0.10
		keyword	dynamik	4528.00	**0.0013**	0.51
	num_correct	normal	keyword	3551.00	0.9418	0.08
		normal	dynamik	2993.00	0.0811	-0.23
		keyword	dynamik	3028.50	0.1041	-0.29
	mental_demand	normal	keyword	3329.00	0.5210	-0.08
		normal	dynamik	3683.50	0.6168	0.13
		keyword	dynamik	3827.50	0.3331	0.21
	physical_demand	normal	keyword	3019.00	0.0989	-0.26
		normal	dynamik	3567.50	0.8987	0.02
		keyword	dynamik	4077.00	0.0742	0.28
	temporal_demand	normal	keyword	2964.50	0.0686	-0.27
		normal	dynamik	3277.50	0.4190	-0.14
		keyword	dynamik	3801.00	0.3793	0.13
	performance	normal	keyword	3642.00	0.7134	0.06
		normal	dynamik	4409.00	**0.0044**	0.44
		keyword	dynamik	4278.00	**0.0155**	0.37
	effort	normal	keyword	3320.50	0.5007	-0.11
		normal	dynamik	4250.50	**0.0195**	0.39
		keyword	dynamik	4427.50	**0.0037**	0.49
	frustration	normal	keyword	2887.50	**0.0390**	-0.32
		normal	dynamik	3478.00	0.8732	-0.03
		keyword	dynamik	4094.50	0.0682	0.29

Table 10: Statistical results by each condition among participants whose Pre-test scores are 7 or lower. For the values of comprehension, engagement, and readability are 0:Not at all – 7:Perfect on Likert Scale. The value of num correct ranges from 0 to 10. For the values of mental demand, physical demand, temporal demand, performance, effort, and frustration are 0:Perfect – 7:Worst on Likert Scale.

Group	Metric	Condition 1	Condition 2	U-statistic	p-value	Cohen's d
Pre-test score ≤ 7	comprehension	normal	keyword	383.00	0.3137	-0.31
		normal	dynamik	626.00	**0.0074**	0.67
		keyword	dynamik	653.00	**0.0022**	0.87
	engagement	normal	keyword	337.00	0.0888	-0.50
		normal	dynamik	575.00	0.0597	0.45
		keyword	dynamik	669.00	**0.0010**	0.91
	readability	normal	keyword	224.00	**0.0007**	-0.98
		normal	dynamik	517.00	0.3156	0.23
		keyword	dynamik	711.50	**0.0001**	1.19
	num_correct	normal	keyword	564.50	0.0870	0.46
		normal	dynamik	365.50	0.2023	-0.32
		keyword	dynamik	268.50	**0.0064**	-0.74
	mental_demand	normal	keyword	337.50	0.0862	-0.35
		normal	dynamik	522.00	0.2792	0.41
		keyword	dynamik	599.00	**0.0250**	0.67
	physical_demand	normal	keyword	408.00	0.5216	-0.19
		normal	dynamik	481.50	0.6340	0.07
		keyword	dynamik	521.00	0.2767	0.24
	temporal_demand	normal	keyword	325.00	0.0565	-0.48
		normal	dynamik	415.50	0.6042	-0.17
		keyword	dynamik	525.50	0.2556	0.28
	performance	normal	keyword	367.50	0.2125	-0.28
		normal	dynamik	604.50	**0.0202**	0.66
		keyword	dynamik	642.00	**0.0040**	0.80
	effort	normal	keyword	356.50	0.1585	-0.33
		normal	dynamik	610.00	**0.0157**	0.72
		keyword	dynamik	678.50	**0.0006**	1.01
	frustration	normal	keyword	254.00	**0.0032**	-0.71
		normal	dynamik	480.00	0.6572	0.16
		keyword	dynamik	632.50	**0.0061**	0.77

Table 11: Statistical results by each condition among participants whose Pre-test scores are higher than seven. For the values of comprehension, engagement, and readability are 0:Not at all – 7:Perfect on Likert Scale. The value of num correct ranges from 0 to 10. For the values of mental demand, physical demand, temporal demand, performance, effort, and frustration are 0:Perfect – 7:Worst on Likert Scale.

Group	Metric	Condition 1	Condition 2	U-statistic	p-value	Cohen's d
Pre-test Score >7	comprehension	normal	keyword	1387.50	0.6556	-0.08
		normal	dynamik	1633.00	0.2585	0.28
		keyword	dynamik	1703.50	0.1141	0.37
	engagement	normal	keyword	1365.50	0.5513	-0.13
		normal	dynamik	1459.00	0.9974	0.05
		keyword	dynamik	1557.00	0.5255	0.18
	readability	normal	keyword	1098.50	**0.0233**	-0.49
		normal	dynamik	1252.00	0.1935	-0.30
		keyword	dynamik	1603.00	0.3642	0.18
	num_correct	normal	keyword	1265.50	0.2245	-0.19
		normal	dynamik	1269.50	0.2340	-0.17
		keyword	dynamik	1455.50	0.9899	0.01
	mental_demand	normal	keyword	1500.50	0.7927	0.03
		normal	dynamik	1467.50	0.9550	0.01
		keyword	dynamik	1394.00	0.6902	-0.02
	physical_demand	normal	keyword	1219.00	0.1320	-0.29
		normal	dynamik	1414.00	0.7825	0.00
		keyword	dynamik	1662.50	0.1987	0.30
	temporal_demand	normal	keyword	1295.00	0.3102	-0.19
		normal	dynamik	1358.00	0.5345	-0.13
		keyword	dynamik	1512.00	0.7383	0.07
	performance	normal	keyword	1613.00	0.3328	0.23
		normal	dynamik	1712.00	0.1115	0.34
		keyword	dynamik	1576.00	0.4597	0.13
	effort	normal	keyword	1490.00	0.8426	0.00
		normal	dynamik	1642.00	0.2502	0.24
		keyword	dynamik	1624.50	0.2989	0.24
	frustration	normal	keyword	1352.50	0.5113	-0.15
		normal	dynamik	1365.50	0.5653	-0.13
		keyword	dynamik	1475.50	0.9154	0.02

Table 12: Statistical results by each condition among non-native English-speaking participants. For the values of comprehension, engagement, and readability are 0:Not at all – 7:Perfect on Likert Scale. The value of num correct ranges from 0 to 10. For the values of mental demand, physical demand, temporal demand, performance, effort, and frustration are 0:Perfect – 7:Worst on Likert Scale.

Group	Metric	Condition 1	Condition 2	U-statistic	p-value	Cohen's d
English Non-Native	comprehension	normal	keyword	2150.00	0.4698	-0.15
		normal	dynamik	2961.00	**0.0033**	0.52
		keyword	dynamik	3067.50	**0.0007**	0.65
	engagement	normal	keyword	2036.50	0.2183	-0.25
		normal	dynamik	2631.50	0.1529	0.25
		keyword	dynamik	2909.50	**0.0074**	0.50
	readability	normal	keyword	1464.00	**0.0002**	-0.69
		normal	dynamik	2396.00	0.7084	0.02
		keyword	dynamik	3166.00	**0.0002**	0.67
	num_correct	normal	keyword	2394.50	0.7154	0.12
		normal	dynamik	2060.00	0.2614	-0.15
		keyword	dynamik	2008.50	0.1773	-0.25
	mental_demand	normal	keyword	2280.00	0.8889	0.02
		normal	dynamik	2578.00	0.2405	0.27
		keyword	dynamik	2569.00	0.2549	0.25
	physical_demand	normal	keyword	2021.50	0.1949	-0.23
		normal	dynamik	2305.00	0.9767	0.01
		keyword	dynamik	2604.50	0.1893	0.24
	temporal_demand	normal	keyword	1968.50	0.1275	-0.27
		normal	dynamik	2077.00	0.2974	-0.20
		keyword	dynamik	2406.50	0.6774	0.07
	performance	normal	keyword	2486.50	0.4390	0.13
		normal	dynamik	2921.00	**0.0070**	0.47
		keyword	dynamik	2735.00	**0.0612**	0.32
	effort	normal	keyword	2049.00	0.2415	-0.20
		normal	dynamik	2824.00	**0.0231**	0.41
		keyword	dynamik	3034.00	**0.0014**	0.59
	frustration	normal	keyword	1901.50	0.0698	-0.31
		normal	dynamik	2314.00	0.9947	0.00
		keyword	dynamik	2699.00	0.0876	0.30

Table 13: Statistical results by each condition among native English-speaking participants. For the values of comprehension, engagement, and readability are 0:Not at all – 7:Perfect on Likert Scale. The value of num correct ranges from 0 to 10. For the values of mental demand, physical demand, temporal demand, performance, effort, and frustration are 0:Perfect – 7:Worst on Likert Scale.

Group	Metric	Condition 1	Condition 2	U-statistic	p-value	Cohen's d
English Native	comprehension	normal	keyword	120.00	0.7663	-0.24
		normal	dynamik	127.00	0.9839	-0.05
		keyword	dynamik	133.50	0.8418	0.18
	engagement	normal	keyword	115.00	0.6093	-0.33
		normal	dynamik	135.00	0.7865	-0.07
		keyword	dynamik	144.50	0.5225	0.22
	readability	normal	keyword	100.00	0.2816	-0.47
		normal	dynamik	87.50	0.1184	-0.63
		keyword	dynamik	118.50	0.7299	-0.07
	num_correct	normal	keyword	112.00	0.5468	-0.07
		normal	dynamik	85.00	0.0944	-0.66
		keyword	dynamik	106.00	0.3933	-0.48
	mental_demand	normal	keyword	92.00	0.1660	-0.47
		normal	dynamik	96.00	0.2169	-0.40
		keyword	dynamik	127.00	0.9843	0.04
	physical_demand	normal	keyword	99.00	0.2742	-0.39
		normal	dynamik	135.00	0.8013	0.07
		keyword	dynamik	160.50	0.2192	0.44
	temporal_demand	normal	keyword	107.00	0.4198	-0.34
		normal	dynamik	134.00	0.8296	0.08
		keyword	dynamik	154.00	0.3162	0.42
	performance	normal	keyword	108.00	0.4547	-0.19
		normal	dynamik	147.00	0.4740	0.37
		keyword	dynamik	171.00	0.1020	0.63
	effort	normal	keyword	149.50	0.4123	0.25
		normal	dynamik	150.50	0.3891	0.38
		keyword	dynamik	131.50	0.9077	0.13
	frustration	normal	keyword	104.00	0.3485	-0.39
		normal	dynamik	114.50	0.6091	-0.15
		keyword	dynamik	141.00	0.6240	0.25

Table 14: NASA-TLX Scores by Participant Group (0:Low – 7:High)

Metric	Statistic	Participant Group				
		All Participants	Pre-test Score ≤ 7	Pre-test Score >7	English Non-Native	English Native
Mental Demand	Mean	3.11	2.98	3.18	3.05	3.35
	Median	3.0	3.0	4.0	3.0	4.0
	Range	0.0–6.0	0.0–6.0	0.0–6.0	0.0–6.0	0.0–5.0
Physical Demand	Mean	1.79	1.94	1.70	1.71	2.12
	Median	1.0	2.0	1.0	1.0	2.0
	Range	0.0–5.0	0.0–5.0	0.0–5.0	0.0–5.0	0.0–5.0
Temporal Demand	Mean	2.89	2.92	2.88	2.75	3.50
	Median	3.0	3.0	3.0	3.0	4.0
	Range	0.0–6.0	0.0–6.0	0.0–6.0	0.0–6.0	0.0–5.0
Performance	Mean	2.59	2.69	2.54	2.71	2.10
	Median	2.0	3.0	2.0	3.0	2.0
	Range	0.0–6.0	0.0–6.0	0.0–6.0	0.0–6.0	0.0–6.0
Effort	Mean	2.90	2.72	2.99	2.73	3.60
	Median	3.0	3.0	3.0	3.0	4.0
	Range	0.0–6.0	0.0–6.0	0.0–6.0	0.0–6.0	0.0–6.0
Frustration	Mean	2.59	2.93	2.40	2.76	1.88
	Median	3.0	3.0	3.0	3.0	2.5
	Range	0.0–6.0	0.0–6.0	0.0–6.0	0.0–6.0	0.0–6.0

Table 15: Other Metrics by Participant Group (As for Comprehension, Engagement, Readability, 0:Not at all – 7:Perfect on Likert Scale. Num Correct ranges from 0 to 10.)

Metric	Statistic	Participant Group				
		All Participants	TOEFL ≤ 7	TOEFL >7	English Non-Native	English Native
Num Correct	Mean	3.94	3.51	4.18	3.85	4.34
	Median	4.0	3.0	4.0	4.0	5.0
	Range	0.0–7.0	0.0–7.0	0.0–7.0	0.0–7.0	1.0–7.0
Comprehension	Mean	3.71	3.35	3.91	3.59	4.23
	Median	4.0	3.0	4.0	4.0	4.0
	Range	0.0–6.0	0.0–6.0	0.0–6.0	0.0–6.0	0.0–6.0
Engagement	Mean	3.57	3.21	3.77	3.46	4.03
	Median	4.0	3.0	4.0	4.0	4.0
	Range	0.0–6.0	0.0–6.0	0.0–6.0	0.0–6.0	0.0–6.0
Readability	Mean	3.88	3.52	4.09	3.78	4.28
	Median	4.0	4.0	4.0	4.0	4.0
	Range	0.0–6.0	0.0–6.0	0.0–6.0	0.0–6.0	1.0–6.0

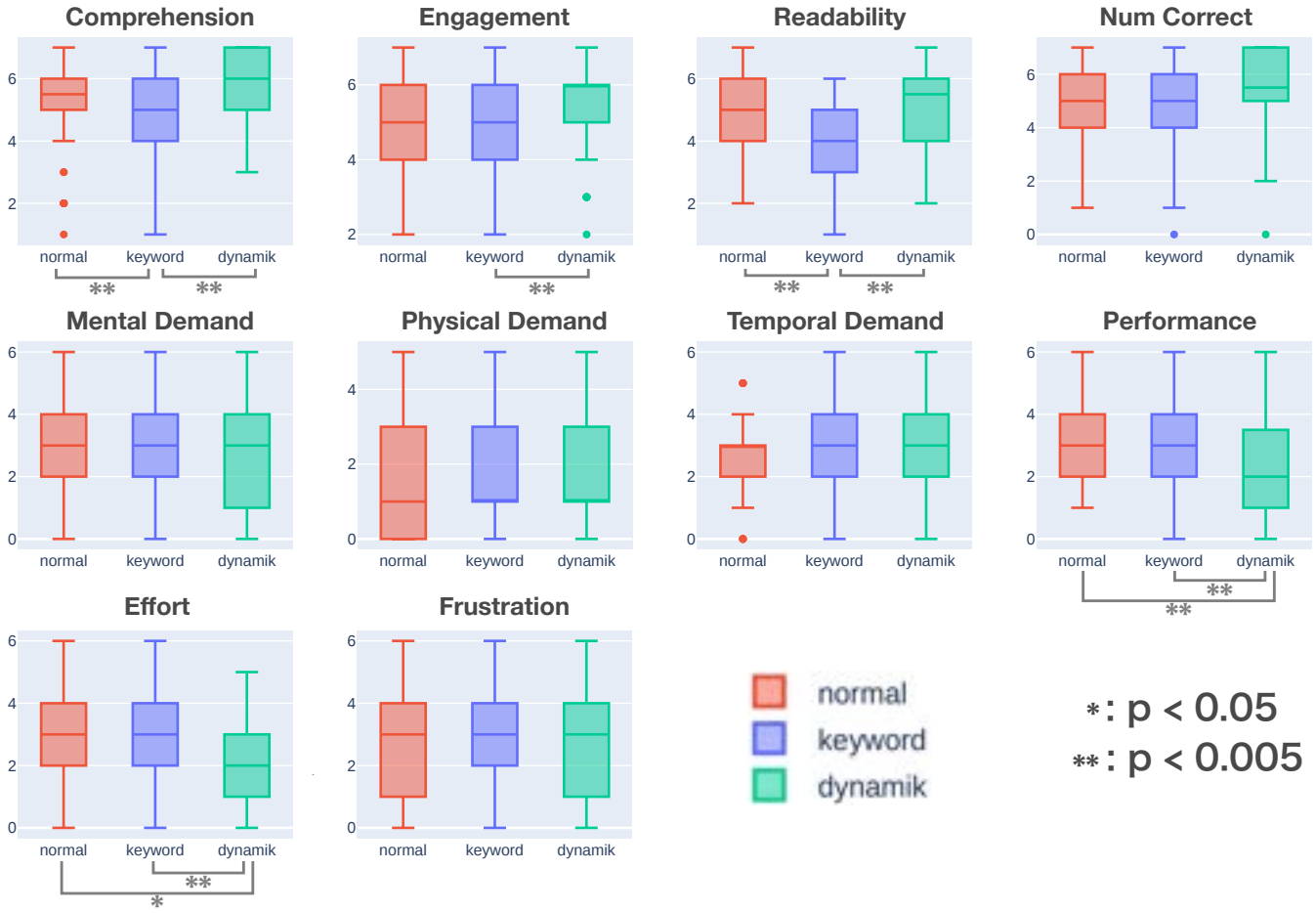


Figure 10: Workload and each item on the questionnaire of non-native English-speaking participants.

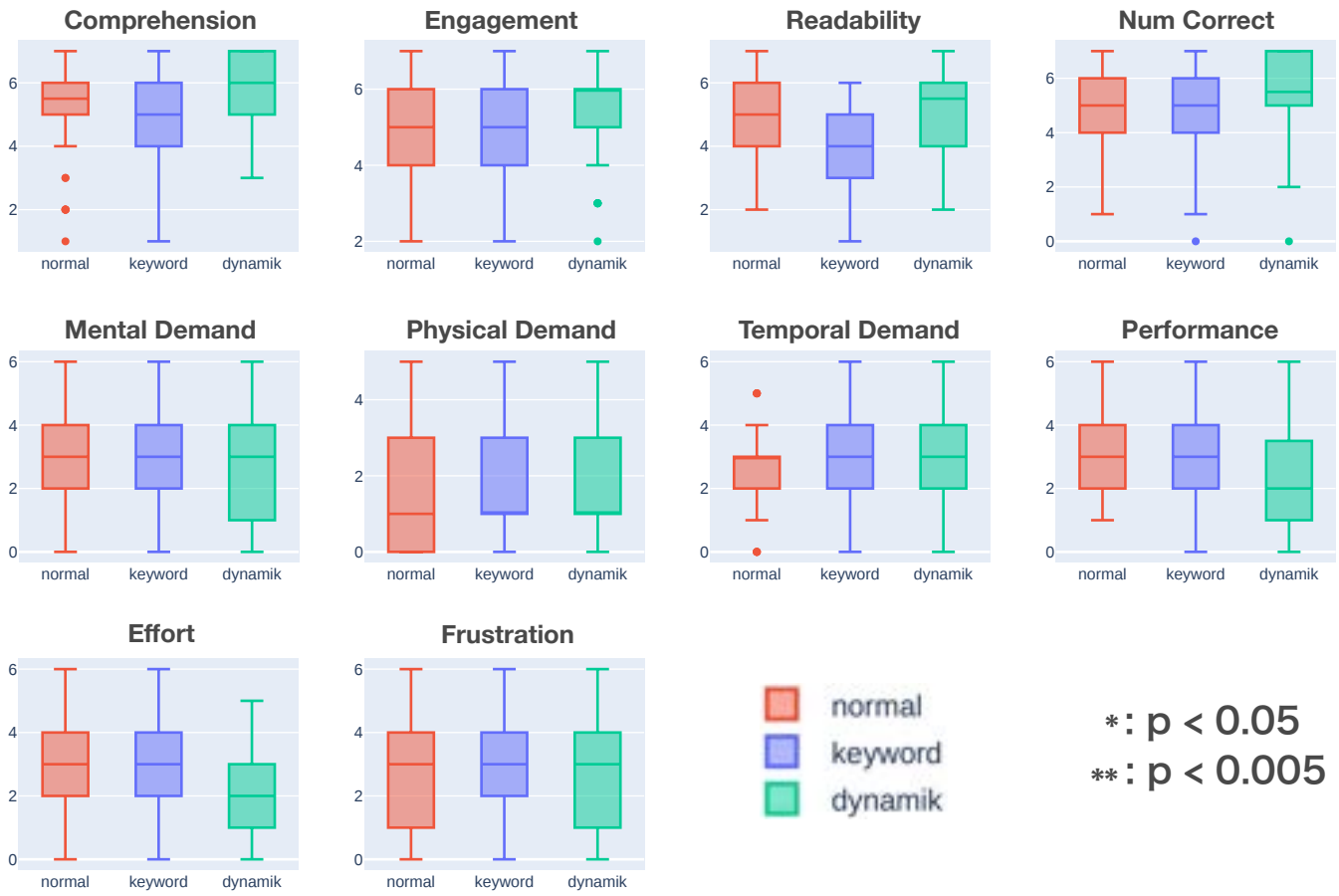


Figure 11: Workload and each item on the questionnaire of native English-speaking participants.