Useful but Distracting: Viewer Experience with Keyword Highlights and Time-Synchronization in Captions for Language Learning

Henrike Weingärtner
LMU Munich
Munich, Germany
henrike.weingaertner@ifi.lmu.de

Maximiliane Windl
LMU Munich
Munich, Germany
Munich Center for Machine Learning (MCML)
Munich, Germany
maximiliane.windl@ifi.lmu.de

Lewis L. Chuang
Department of Humans and Technology
TU Chemnitz
Chemnitz, Germany
lewis.chuang@phil.tu-chemnitz.de

Fiona Draxler
University of Mannheim
Mannheim, Germany
fiona.draxler@uni-mannheim.de

Figure 1: The four selected caption designs. (A) Standard captions, (B) full captions with keyword highlights, (C) timed keyword-only captions, (D), full captions with timed keyword highlights, where each keyword is highlighted when it is pronounced.

MUM '24, December 1–4, 2024, Stockholm, Sweden

https://doi.org/10.1145/3701571.3701574

Abstract

Captions are a valuable scaffold for language learners, aiding comprehension and vocabulary acquisition. Past work has proposed enhancements such as keyword highlights for increased learning gains. However, little is known about learners' experience with enhanced captions, although this is critical for adoption in everyday life. We conducted a survey and focus group to elicit learner preferences and requirements and implemented a processing pipeline

for enhanced captions with keyword highlights, time-synchronized keyword highlights, and keyword captions. A subsequent online study (n = 66) showed that time-synchronized keyword highlights were the preferred design for learning but were perceived as too distracting to replace standard captions in everyday viewing scenarios. We conclude that keyword highlights and time-synchronization are suitable for integrating learning into an entertaining everyday-life activity, but the design should be optimized to provide a more seamless experience.

CCS Concepts

• **Applied computing** → **E-learning**; • **Computing methodologies** → *Speech recognition*; • **Human-centered computing** → User studies.

Keywords

language learning, captions, video, speech alignment

ACM Reference Format:

Henrike Weingärtner, Maximiliane Windl, Lewis L. Chuang, and Fiona Draxler. 2024. Useful but Distracting: Viewer Experience with Keyword Highlights and Time-Synchronization in Captions for Language Learning. In *International Conference on Mobile and Ubiquitous Multimedia (MUM '24), December 1–4, 2024, Stockholm, Sweden.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3701571.3701574

1 Introduction

With streaming services and online video platforms, language learners have gained access to potentially unlimited content. Thanks to foreign-language audio and captions, they can improve their skills while watching their favorite show. However, captions on streaming platforms and other media providers are primarily designed for comprehension, not for engaging learners. For example, they include elements that are essential for deaf or hard-of-hearing viewers but may distract language learners, e.g. textual sound descriptions such as [footsteps approaching] or [*Dancing Queen* playing on the radio]. Thus, optimizing captions with the learners' viewing experience in mind could motivate them to watch foreign-language media with captions in everyday life, increasing their foreign-language exposure.

Past work has already explored modifications of captions such as keyword captions [14, 34], captions including keyword translations [18], or interactive support based on eye tracking [13]. Several studies show increased learning gains for such enhanced captions [5, 18, 46]. So far, it remains unclear whether learners actually like these enhanced captions and would be willing to use them in everyday viewing scenarios.

In this paper, we applied a user-centered design process (cf. Figure 2) to implement closed captions enhanced for language learning and evaluate user experience and perceived usefulness as major factors influencing long-term adoption [49]. We target learners at a medium-to-high target language proficiency because we expect them to benefit from captioned video without feeling overwhelmed. As a first step, we identified learner needs in a focus group and an initial survey. Based on related work and our insights from the survey and focus group, we implemented a processing system for three enhanced caption types: (1) captions consisting only of time-synchronized keywords, (2) captions with keyword highlights, and (3) captions with time-synchronized keyword highlights. Words were considered keywords if they were not included in an English-language CEFR[1] A1-B1 corpus. As a baseline design, we added standard full captions. We compared the viewing experience in terms of hedonic and pragmatic quality as well as perceived understanding with these four caption types in an online survey using excerpts from the movie Marriage Story. We found that (time-synchronized) *Keyword Highlights* captions outperformed *Standard Captions* and *Timed Keywords* questions with regards to hedonic qualities and scored almost as high as *Standard Captions* on pragmatic qualities and perceived comprehension. However, the distractions caused by the highlights meant that a majority of users still preferred standard captions, except when they explicitly aimed at learning.

In sum, we contribute (1) a choice of three enhanced caption types that are promising from a user experience perspective, (2) a comparative evaluation of these caption types with regard to user experience and perceived comprehension, and (3) a discussion of implications for embedding captioned viewing in everyday life to support language learning.

2 Related Work

Foreign-language videos, be it movies or TV shows, are a great tool for language learning: they immerse learners in a foreign culture [16], enable comprehension practice [39], and promote vocabulary learning [41, 42]. Generally speaking, videos provide exposure to authentic language, which is beneficial for language acquisition according to Krashen's input hypothesis [4, 19]. This section summarizes how learning can be supported through captions and subtitles. Like Vanderplank [48, p. 9], we use the term *captions* to refer to intralingual or same-language subtitles and *subtitles* to refer to interlingual or foreign-language subtitles. We discuss advanced caption design concepts that utilize the flexibility of current-day media players to optimize the viewing and learning experience for different target groups and briefly address technological prerequisites.

2.1 Captions and Subtitles in Language Learning

Captions and subtitles foster language learning through improved content comprehension [2], listening comprehension [14], vocabulary acquisition [15, 35], and to some extent, also grammar learning [5]. For example, a study on content and listening comprehension showed that students who watch videos with subtitles or captions write better summaries than students without captions [27]. Similarly, learners provided with captions achieved higher scores in comprehension questions than those without [14]. In terms of vocabulary learning, studies have observed both recall and recognition improvements when watching videos with captions or subtitles [15, 35]. How many words a viewer learns depends on factors such as the words' imagery potential and whether the words sound similar to first-language words [40]. Interestingly, Chen et al. [3] found larger vocabulary gains for more proficient students. Studies on grammar learning through subtitles are scarce overall. For example,

[1]European Reference Scale; https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

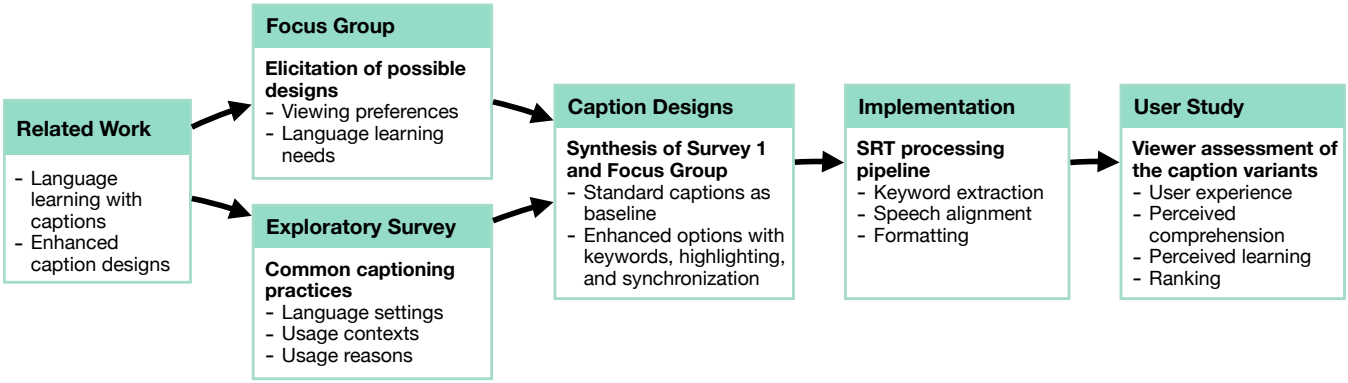

Figure 2: The steps of the design and evaluation process employed in this paper.

Cintrón-Valentín et al. [5] found positive effects of textual enhancements in captions, but only for some of the enhanced structures, while a study with children by Lommel et al. [25] showed no effects on grammar learning. One important aspect to consider for learning success is cognitive load. On the one hand, the combination of multiple modalities—the associations of images, written, and spoken words—supports dual coding [8, 29] and can lead to a greater depth of processing [9]. On the other hand, subtitles add an additional information channel that viewers need to process, and this can potentially cause a high cognitive load. Accordingly, a study by Taylor [45] showed that many first-year learners found captions distracting and that adding captions impacted their listening comprehension. However, this was not the case for third-year learners who already had more language exposure. Similarly, an eye tracking and EEG study by Kruger et al. [20] showed that despite the verbal redundancy effect, the risk of cognitive overload caused by captions was low. Therefore, our target group in this work is also learners with a medium to high target language proficiency. An outlook on additional aspects, such as the suitability of different video genres and recent work on learner strategies, is provided in the literature reviews by Vanderplank [48] and Montero Perez [32].

The cited literature above includes work on intralingual captions and interlingual subtitles. In fact, research so far has not shown conclusive evidence in favor of one or the other [28, 36]. Unsurprisingly, subtitles are particularly helpful for content comprehension of novice learners [2]. However, another study found that learners watching a video with Scottish or Australian accents and English captions were better at understanding and repeating words than a Dutch subtitle group [31]. Regarding vocabulary learning, a 7-week study by Frumuselu et al. [12] indicated that both novice and advanced learners perform better when using captions. Moreover, Markham et al. [27] suggest advancing from subtitles to captions to no captions on subsequent viewings as a beneficial strategy.

In sum, the decision to use captions or subtitles depends on the learner's goal and context. In this work, we focus on intralingual captions because of their widespread availability, or as Vanderplank [48] put it:

> *We are [...] fortunate that those with a disability have provided us, who are merely 'hard-of-listening' in a foreign language, with a wonderful resource not only for making films and TV programmes accessible to us but for helping us improve our reading, listening, and speaking skills.*

2.2 Enhanced and Interactive Caption Design

Above, we discussed standard full-text subtitles and captions. However, with current-day media players, loading new subtitle files has become very easy. This opens up new possibilities for static, adaptive, or even interactive subtitle and caption variants. For example, static subtitle adaptations include captions that only show keywords [14, 15, 33, 47] or highlight target word [26]. Both of these approaches can benefit learning by increasing the focus on target words or reducing distractions. However, this does not necessarily increase recall of target words in comparison to default captions [33, 47]. Other proposed methods add keyword translations, similar to text glosses [43, 46]. However, a major challenge with keyword or highlight captions is the selection of appropriate keywords, as it is difficult to assess what learners already know. A common approach is to select words based on their frequency in corpora, such as the BNC/COCA lists for English [38]. Guillory [14] had experts choose the words that were deemed most difficult. As a further adaptation, Kurzhals et al. [22] proposed speaker-following subtitles, which clearly mark the connection between speaker and dialog content and, thus, may reduce eye strain by reducing saccade length [11, 22]. However, this approach requires advanced preprocessing. Wang and Pellicer-Sánchez [50] investigated the effectiveness of bilingual subtitles compared to captions, subtitles, and no subtitles using an eye-tracking study. Thereby they found that while bilingual captions lead to a higher meaning recognition, they can also be distracting as users tend to spend more time reading the translations than the new words in the target language. Mirzaei et al. [30] investigated synchronizing speech signal and keyword captions and found short-term enhancements on subsequent viewing of non-captioned videos in comparison to full captions and no captions. Finally, several projects and studies have explored interactive subtitles. For example, Kovacs and Miller [18] enhanced captions with features for interactive vocabulary lookup, line translation, video navigation, and transcription to an alphabet familiar to the learner. This increased vocabulary learning in comparison to dual-language subtitles. However, the information-dense subtitles led to viewing

times between 10 and 12 minutes for 5-minute videos, thus substantially changing the experience from linear viewing. In addition, Zhu et al. [51] designed a dictionary where entries are enriched with captioned video clips, including target word highlights and translations, resulting in higher vocabulary retention than with a traditional dictionary. Commercial platforms such as FluentU[2], LingoPie[3], and Language Reactor[4] also provide interactive captions for language learning and promote this as an enjoyable way of learning. Since our objective is to integrate learning using captions into everyday viewing experiences, we do not include interactive elements that may shift the focus toward learning and consequently impact entertainment and long-term motivation. Thus, we apply a static approach with preprocessed subtitle files.

2.3 Subtitle Files and Subtitle Processing

Srt files are well-suited for simple adaptations because they are human-readable and supported by common media players such as VLC and can even be activated on top of browser-based Netflix and other video-on-demand players with extensions such as Substital[5]. However, they also come with several drawbacks. Notably, srt files are often unofficially distributed, are more easily available for blockbusters than arthouse movies, and frequently contain mistakes. In addition, the ideal timing can differ depending on the associated media type. For example, there may be additional opening credits in a BluRay version that are not shown by a video-on-demand provider, and this delays the timing of the BluRay subtitles, requiring manual synchronization or a tool such as Laiola Guimarães et al.'s framework [24].

3 Survey on Caption Usage

As the first pointer towards favored caption designs for language learners, we surveyed 61 people on their current caption or subtitling preferences and usage habits. Specifically, we asked them how often they use captions or subtitles, what languages they set them to, and how much they like watching video material with captions.

3.1 Participants

The 61 respondents were recruited via university mailing lists. They were between 17 and 65 years old ($M = 27.0, SD = 8.9$ years). Thirty-nine participants identified as female, 20 as male, one as diverse, and one did not disclose their gender. Fifty-nine participants were native German speakers, and two were native Russian speakers. Five participants listed a second native language (Italian, Russian, Farsi, or Spanish). The survey was conducted in German. Note that we used the German term "Untertitel", which encapsulates both captions and subtitles. We incentivized participation with a raffle of 20€ vouchers (one per ten participants).

3.2 Survey Results

The survey results revealed diverse subtitling and caption habits and preferences. This was already apparent from the caption usage within the last thirty days: 16% of the participants reported never using captions, 26% used them a few times per month, and 57% used captions weekly or daily. These results align with a 2022 US survey, where 50% of respondents (70% of Gen Z respondents) said they watch content with subtitles or captions most of the time [37]. A majority of respondents stated that they used captions in the video language (74%) or subtitles in their native language (45%; multiple responses possible). 25% also set subtitles to a third language, for example, when the available options are limited or when they are watching with someone else. The primary reasons for activating captions were insufficient language skills (74%), distractions caused by a noisy environment (67%), a low video volume (51%), other people needing subtitles (51%), a fast rate of speech (46%), dialects (43%), difficult words (38%), for language learning (5%), unintelligible pronunciation (3%), or when watching without sound (3%). Responding to the phrase "I like subtitles", 46% of participants agreed with the statement, 23% reported a neutral feeling, and 31% disagreed.

Overall, the survey highlights that subtitles and captions are frequently used. Most participants in our sample activate subtitles for better comprehension, whereas only a few intentionally do so for language learning. This points to an opportunity to increase the motivation to learn by adapting the caption design to improve the language learning experience.

4 Focus Group on Preferred and Envisioned Caption Designs

We conducted an online focus group with six participants to discuss how captions can be adapted to cater to the needs and viewing experience of language learners. First, we presented and discussed current caption solutions beyond traditional closed captioning. Then, we asked our participants to develop their own ideas.

The participants (three male and three female) were between 20 and 30 years old. They were all native German speakers and had learned English in school.

4.1 Procedure

After an introduction round, we showed the participants short video clips with caption designs from or inspired by prior work. We asked them to discuss the concepts in light of their usefulness for language learning. The first five clips were shown in one go; the last three were presented one after the other whenever the conversation had come to a hold. Overall, we showed eight subtitle variants as a basis for discussion. These covered a range of novel features such as translations, highlights, and dynamic positioning:

(1) Captions with translations and explanations for individual words as in Zhu et al. [51]
(2) Captions with translations of words on hover as in Kovacs and Miller [18]
(3) Captions with keyword highlighting and an additional text box with keywords and their translations as in Ma et al. [26]
(4) A modified version of the latter without highlights and translations
(5) Another modified version of Ma et al. [26] without the standard captions
(6) Captions with translations in parentheses as in Sakunkoo and Sakunkoo [43]

[2]https://www.fluentu.com, last accessed 2024-08-15
[3]https://lingopie.com, last accessed 2024-08-15
[4]https://www.languagereactor.com, last accessed 2024-08-15
[5]https://chrome.google.com/webstore/detail/substital-add-subtitles-t/kkkbiiikppgjdiebcabomlbidfodipjg, last accessed 2024-08-15

(7) Displaying captions next to the person speaking as in Kurzhals et al. [22]
(8) Rather than spoken words, the last variant presented in-place object labels and translations. This variant showcased caption use beyond dialogues.

Following the discussion, the participants engaged in an ideation activity using the 6-3-5 brainwriting method[6] on a collaborative board with digital sticky notes. In the end, they shared and discussed their ideas with the group. The focus group was conducted in German.

4.2 Findings of the Focus Group

The discussion in the focus group highlighted the importance of avoiding disruptions and considering cognitive demands while catering to situation-dependent information needs. The ideation phase provided a starting point for further exploration of adaptations and novel caption designs.

Disruptions and Cognitive Load. The participants identified attention switches caused by the caption design as potential sources of disruption. They were also afraid that overloaded designs would make them miss parts of the movie. Our participants considered this particularly critical for caption variations that included translations and redundant or non-essential information. Specifically, they emphasized that native-language translations immediately and automatically attract attention, limiting the resources available for the original captions and the scene content. In addition, they found translations particularly distracting when the original caption and the translation used different alphabets. When translations were to be displayed, participants preferred them to be positioned under the original word rather than in a separate keyword box to minimize lookup times. Participants also said that only words that are actually pronounced should be displayed. Even for genres with less focus on narrative and conversations, such as documentaries, they considered object label captions (keyword variant 8) not helpful because of the already inherent factual learning focus. In sum, our participants were afraid they could not focus on more than one thing at a time.

Situation- and User-Dependent Information. Participants noted that the requirements for captions depend on individual and situational factors such as the language level and the speakers' dialect or rate of speaking. For example, they positively commented on the captions that moved along with speakers, in particular for speakers with strong accents or dialects. However, they felt that the display time might be too short for following fast speakers. They also found the idea of keyword captions interesting. Keywords reduce the overall information load and can target words that are specifically helpful for learners of a given language level. For translated keyword captions, participants feared that they might not always be able to recognize them when they are pronounced. Finally, they also discussed the timing of words so that they appear the moment they are pronounced. This way, viewers could immediately connect words with their pronunciation.

[6]https://en.wikipedia.org/wiki/6-3-5_Brainwriting

Extensions and Novel Ideas. Based on the discussed caption designs and their own experience, the participants came up with novel ideas and extensions of the presented caption variants. These ideas can be grouped into concepts that focus either on comprehension or learning. For better comprehension, suggestions include selective captioning of characters that speak dialects or are hard to follow. Similarly, captions could highlight technical terms or words that occur particularly infrequently and are, thus, more likely to be unknown. For learning, participants felt that it might be helpful to add or highlight homonyms, typical idioms, dialectal differences, and/or words without direct translations. In an interactive system, translations could be shown on request. Grammatical support could be provided, e.g., by coloring different tenses, endings, word boundaries, or functions of words. Moreover, the level of detail should be adaptable to match the viewers' language level.

5 Final Caption Designs and Hypotheses

The user-centered design process including the initial survey and focus group motivated our final selection of caption designs as detailed below. We made sure to include both traditional designs and enhanced suggestions such as highlighting and keywords. We opted against translations to reduce mental load (cf. section 4.2). Before comparing the enhanced caption designs in a user study, we derive hypotheses regarding the expected effect on user experience, perceived comprehension, perceived learning, and vocabulary recall.

5.1 Selected Caption Designs

Based on past literature, the focus group, and the survey, we finally selected the following four caption designs that vary between focusing on target words through keywords and providing context through full captions (cf. Figure 1):

(1) ***Standard Captions.*** This variant represents the state of the art and serves as a baseline.
(2) ***Keyword Highlights.*** This variant shows full captions with keyword highlights and is based on Ma et al. [26]. However, we do not show translations of the words because the participants in the focus group considered translations distracting.
(3) ***Timed Keywords.*** This caption type shows keywords at the exact time they are spoken, while all other words are removed. The idea is based on Mirzaei et al. [30], who proposed timed keyword captions as a means to focus on vocabulary learning without the distraction caused by the full transcript. Our focus group also confirmed the potential of time synchronization.
(4) ***Timed Keyword Highlights.*** This is a full-caption variant with timed keyword highlights. Thus, it is a hybrid of *Keyword Highlights* and *Timed Keywords*. It was introduced to guide the viewers' attention while still providing context. With this variant, we aim to compensate for the potential mismatch between keyword selection and learner knowledge.

5.2 Hypotheses

We derive the following hypotheses concerning measures for user experience (UX), perceived comprehension and learning, and vocabulary recall. Assessing UX and perceived comprehension helps us understand what type of captions learners are potentially willing to use in everyday life. We also added vocabulary recall to position the effectiveness of our designs in relation to prior work, but this was not our primary focus. Hypotheses are based on related work, the focus group, and the survey.

- H1a: The PRAGMATIC QUALITY is rated highest for *Standard Captions*. We expect this as viewers know this variant and feel most comfortable using it.
- H1b: The HEDONIC QUALITY is rated highest for *Timed Keyword Highlights*. We expect this variant to be considered innovative and providing a good balance between context on the overall scene and focus on potentially challenging aspects.
- H2: *Timed Keyword Highlights* and *Keyword Highlights* achieve the best PERCEIVED COMPREHENSION. Conversely, *Timed Keywords* achieve the lowest perceived comprehension. Again, we assume the focus on potentially challenging aspects to be crucial. Even though Mirzaei et al. [30] stressed the advantage of reducing captions to keywords and reducing reading times, we expect that the lack of context hinders understanding, especially when the keyword selection is not perfectly matched to the viewers' language level.
- H3: *Timed Keyword Highlights* and *Keyword Highlights* fare best for PERCEIVED LEARNING. These are followed by *Timed Keywords* because despite the focus on target words, viewers perceive a lack of context; *Standard Captions* are perceived as least suitable for learning.
- H4: Both highlighted variants and *Timed Keywords* improve VOCABULARY RECALL scores of keywords in comparison to standard *Standard Captions*. As all three enhanced designs put additional focus on keywords, we expect them to attract the viewers' attention.

6 Caption Generation and Video Preparation

We manipulate original srt files by removing non-keywords, adding highlights, or running forced alignment to adjust timestamps. Appendix C gives an overview of the processing pipeline. We use a Python architecture with the pysrt package[7] for working with subtitle files. For all variants, the first step is the detection of keywords to determine what needs to be displayed or removed and what needs to be highlighted. We follow a reverse approach, i.e., we mark a word as a keyword if it does not occur in non-keyword lists. For identifying words at a specific language level, we follow Andrade [1], who analyzed the vocabulary usage of a large number of movies. In particular, we merge the Oxford 5000 list[8] with the BNC/COCA corpus [38] to estimate the language level of word stems and the derived word forms and to remove proper names. When word levels are not uniquely identifiable (e.g., the stem "accept" is considered an A1 word, while "acceptance" is C1), we manually check for false positives. That is, we remove easy and frequent words that are not actually B2+ keywords. We then mark keywords in the subtitles files with HTML font styling. Finally, we run a Gentle[9] server for forced speech alignment. In case a keyword is highlighted for less than 500ms, we extend the display duration by 300ms or until the next caption line is shown.

We also used the script proposed by Andrade [1] to determine suitable scenes. For this, we evenly partitioned the subtitle file into 30 parts and counted B2+ word (keyword) occurrences in each part. We manually extracted scenes from high-keyword partitions and verified that the scenes did not include explicit content. Finally, we prepared all four caption types for the resulting four movie clips of 2-3 minutes, leading to 16 preprocessed caption + video combinations. The video clips contained 24, 30, 39, and 41 keywords, respectively. Because of the higher density of keywords and partially overlapping speech, clips three and four were slightly more difficult than the first two.

7 User Study

To assess the hypotheses introduced in Section 5.2, we conducted a within-subject study with 66 participants. Specifically, we compared the user experience, learning, and perceived comprehension with the four different caption types applied to four scenes from the movie *Marriage Story*, a 2019 movie that follows a couple's divorce. As one of the proposed top 10 movies for "people at C1 level" [1], *Marriage Story* is suitable for our target group of medium- to high-proficiency learners. The movie contains many dialogues, is non-violent overall, and it was easy to select non-explicit scenes with a diverse vocabulary.

7.1 Procedure

The study was implemented as an online survey and could be taken in Spanish or German. Once participants had read the study information and given their consent, we asked them about their experience with subtitles and their prior knowledge of English. We also included a vocabulary pre-test modeled after Nation's Vocabulary Size Test[10]. The pre-test included multiple-choice questions on five keywords from each scene and four distractor items that did not occur in the videos. Participants then watched four movie clips, each with a different condition. Directly after each video, they responded to the UEQ-S [44] in the official Spanish or German version[11]. They rated their comprehension of the content and language and their overall impression of the caption variant. We applied Latin square counterbalancing to vary the order of presentation and the pairing of the movie clip and caption variant, using four of the 16 preprocessed videos for each participant. After the four clips, we asked participants to what extent they had focused on learning, comprehension, and entertainment and asked them to rank the suitability of the caption variants for these goals. The last part of the survey was a vocabulary post-test. Finally, two days later, participants took a second vocabulary post-test to accommodate for initial memory consolidation [10]. We provide a full list of measures in Appendix B. We collected the demographics via Prolific.

[7] https://github.com/byroot/pysrt
[8] https://www.oxfordlearnersdictionaries.com/about/wordlists/oxford3000-5000
[9] https://github.com/lowerquality/gentle
[10] https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests, last accessed 2024-08-15
[11] Translations taken from https://www.ueq-online.org

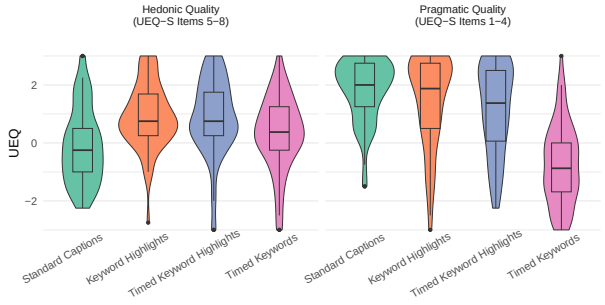

Figure 3: UEQ-S scores of the four caption types compared in our user study.

7.2 Participants

We recruited native Spanish and German speakers that did not live in English-speaking countries via Prolific[12]. Sixty-six participants completed the study. Of these, 23 identified as female and 43 as male. They were between 19 and 60 years old ($M = 32.9$, $SD = 10.8$ years). The 23 German speakers were residents of Germany (17), Austria (5), and Switzerland (1). The 43 Spanish speakers were residents of Mexico (18), Spain (19), Chile (5), and Portugal (1). They self-assessed their English level at A1 (2), A2 (3), B1 (18), B2 (17), C1 (21), or C2 (5) on the CEFR scale. The study took approximately 45 minutes, and participation was compensated with £8.5.

8 Results of the User Study

This section presents the study results with a focus on the participants' experiences and perceptions, following the hypotheses from Section 5.2 and closing with a final ranking and outlook on participants' envisioned designs.

8.1 Analysis

We validate the hypotheses for the four caption types with a repeated-measures ANOVA, with *Standard Captions* serving as the baseline comparison. We apply a Greenhouse-Geisser correction when a Mauchly's test indicates a violation of the sphericity assumption. In case of a significant result ($\alpha < 0.05$), we follow up with pairwise post-hoc tests using a Holm correction and report Cohen's d for effect sizes. We apply non-parametric Friedman tests with Holm-corrected Conover post-hoc tests for questions with a single ordinal scale. All tests are performed with JASP [17]. To illustrate potential explanations of identified trends, we augment the report with exemplary participant statements[13]. For the preferred caption designs, we cluster all available responses and inductively derive general themes.

8.2 User Experience (H1)

As seen in Figure 3, *Keyword Highlights* and *Timed Keyword Highlights* were rated best on the UEQ-S items representing the hedonic quality ($F(2.55, 166.0) = 12.71$, $p < 0.001$, $\eta^2 = 0.16$). Pairwise post-hoc tests show significant differences between almost all conditions: *Standard Captions* fare worse than *Timed Keyword Highlights* ($t = -5.02$, $p < 0.001$, $d = -0.78$), *Keyword Highlights* ($t = -5.31$, $p < 0.001$, $d = -0.82$), but not *Timed Keywords*. *Keyword Highlights* was rated better than *Timed Keywords* ($t = 3.15$, $p < 0.01$, $d = 0.49$), and so was *Timed Keyword Highlights* ($t = 2.85$, $p < 0.05$, $d = 0.44$). There were also significant differences for the pragmatic quality ($F(2.90, 188.5) = 58.59$, $p < 0.001$, $\eta^2 = 0.47$). *Timed Keywords* were clearly outperformed by the three other conditions. Accordingly, pairwise comparisons show that *Timed Keywords* performs significantly worse than *Standard Captions* ($t = 12.13$, $p < 0.001$, $d = 1.84$), *Keyword Highlights* ($t = 10.43$, $p < 0.001$, $d = 1.58$), and *Timed Keyword Highlights* ($t = 8.98$, $p < 0.001$, $d = 1.36$). *Standard Captions* was also considered better than *Timed Keyword Highlights* ($t = 3.15$, $p < 0.01$, $d = 0.48$). The remaining comparisons showed no significant differences.

In H1a, we posited that the pragmatic quality would be rated highest for *Standard Captions*. However, *Timed Keyword Highlights* performed similarly well. As expected in H1b, the hedonic quality was highest for *Timed Keyword Highlights*, although *Keyword Highlights* came close. Thus, time-synchronization was not rated as well as expected.

8.3 Perceived Comprehension of Language and Content (H2)

As shown in Table 1, *Standard Captions*, *Keyword Highlights*, and *Timed Keyword Highlights* were perceived as similarly good for content and language comprehension. *Timed Keywords* was significantly worse than all other conditions (all $p < 0.01$). Nonetheless, all caption types substantially contributed to language and content comprehension, with no median score below 5 (out of 6). This means that as predicted in H3, *Timed Keywords* achieved the lowest perceived comprehension. However, contrary to our expectations, *Standard Captions* was comparable to *Keyword Highlights* and *Timed Keyword Highlights*.

8.4 Perceived Learning (H3)

On the question "I feel that I can learn new words very well with this caption variant," *Timed Keyword Highlights* and *Keyword Highlights* achieved the highest median value (cf. Table 1). Conover post-hoc tests indicated that *Timed Keywords* captions were significantly less suitable for learning than the other three types (all $p \leq 0.01$). There were no significant differences between the other conditions, and H3 cannot be confirmed.

8.5 Vocabulary Recall (H4)

The participants' prior knowledge of the tested vocabulary was high overall. On average, they correctly answered 84.6% of the 24 questions in the vocabulary test before watching the videos, 84.6% in the test right after, and 85.3% in the 2-day delayed post-test. There were no differences in the score changes from before watching the videos to the 2-day delayed post-tests when differentiated by caption type. We observed clear ceiling effects: some participants already knew all the words tested for a condition and could, therefore, not improve their score. In the survey, two people admitted that they looked up words, and several others may have done so. All in all, we cannot confirm H4. We did not identify any differences in the keyword recognition scores.

[12]https://prolific.co
[13]Translated to English if necessary

Table 1: Median and standard deviation showing the agreement to opinion statements for each condition. Responses range from 1 ("I strongly disagree") to 6 ("I strongly agree").

		Standard Captions		*Keyword Highlights*		*Timed Keyword Highlights*		*Timed Keywords*	
	Friedman test	***MD***	***SD***	***MD***	***SD***	***MD***	***SD***	***MD***	***SD***
I understood the language well.	$\chi^2(3) = 25.0, p < 0.001, W = 0.13$	6	1.08	6	1.21	6	1.17	5	1.40
I understood the content well.	$\chi^2(3) = 38.2, p < 0.001, W = 0.19$	6	0.91	6	0.91	6	0.89	5	1.27
I feel that I can learn new words very well with this caption variant.	$\chi^2(3) = 46.9, p < 0.001, W = 0.24$	5	1.45	5	1.27	4	1.54	3	1.62
Viewing the video with this type of caption was agreeable.	$\chi^2(3) = 84.9, p < 0.001, W = 0.43$	5	1.01	4.5	1.40	5	0.89	2	1.50
I can very well imagine using this type of caption.	$\chi^2(3) = 90.3, p < 0.001, W = 0.46$	5	1.21	5	1.64	5	1.78	1.5	1.51

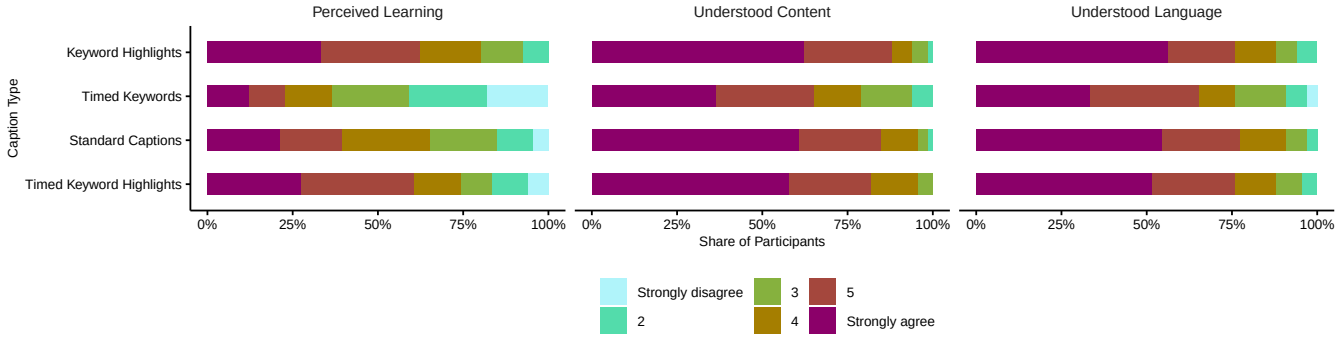

Figure 4: Perceived learning and comprehension with the four caption variants. The full agreement statements were "I feel that I can learn new words very well with this caption variant." (left), "I understood the content well" (center), "I understood the language well." (right), rated on a scale from 1 (Strongly disagree) to 6 (Strongly agree).

8.6 Final Ranking

The assessments above also align with the final ranking of the suitability for comprehension, entertainment, and learning after watching all videos (cf. Figure 5). *Standard Captions* captions were top-ranked for comprehension and entertainment, while *Timed Keyword Highlights* was top-ranked for learning. *Timed Keywords* obtained the lowest overall ranking for all three use cases. This was also reflected in the absolute rating of the caption types: On a scale from 1 to 7, *Standard Captions* best ($MD = 6$, $SD = 1.37$). *Keyword Highlights* ($MD = 6$, $SD = 1.65$) and *Timed Keyword Highlights* ($MD = 6$, $SD = 1.75$) were comparable, and *Timed Keywords* only achieved a median rating of 2 ($SD = 1.78$).

The participants' statements on the caption types give insights into possible reasons for the individual rankings. Notably, *Standard Captions* captions were considered helpful for comprehension because they are *"familiar"* (P49), *"straightforward"* (P48), and *"efficient and non-disruptive"* (P18). P12 described this type as *"Very clear, I understood everything perfectly."* According to P21, they are *"excellent for understanding spoken English in specific contexts."* Typical comments explaining the participants' assessment of the *Keyword Highlights* and *Timed Keyword Highlights* captions show that they were considered helpful but also distracting. For example, for *Keyword Highlights*, P27 noted that *"as long as the video and audio are aligned, this type of viewing captions is agreeable to also learn sentence construction and figures of speech. Sometimes, it distracts from the video because it takes more time to read the full sentences."* Similarly, P10 explained that *"if you want to pay attention to [comprehension and learning], highlighted words distract a bit. I see their use when someone is trying to learn new vocabulary."* P3 felt that *"highlighting some words can make you loose time while reading because the brain will focus on this specific word."* Time-synchronization tended to increase the perceived level of distraction: P17 stated that they *"started to think about which word will turn yellow next"* and P18 added that *"The yellow words can be a bit distracting for*

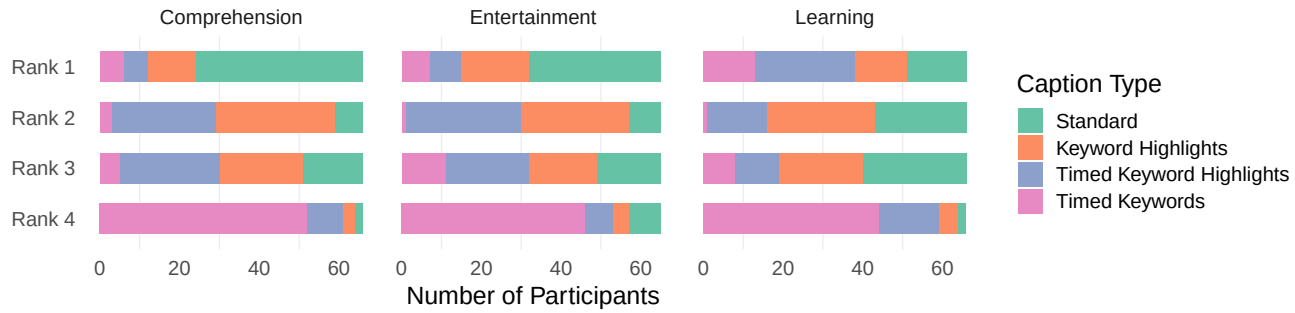

Figure 5: Ranking of the caption types for the purposes *comprehension*, *entertainment*, and *learning*

people that already [know] pretty well the meaning." Similarly, P7 liked seeing the highlights before they were spoken, so *"you can anticipate the focus on the moment where it is mentioned."* On the other hand, P49 found that *Timed Keyword Highlights* captions seemed to *"support you in paying more attention to the plot than with 'normal' subtitles."* The comments also illustrate why some participants felt that *Timed Keywords* captions were not ideal for content and word comprehension. For example, P1 noted that they felt *"distracted"* because this caption type was *"more focused on drawing the attention towards certain words than on helping with the plot."* Moreover, eleven participants explicitly mentioned that they lacked context when they only saw keywords or preferred types that provided full context. For example, P42 said *"The keywords alone do not contribute at all to the understanding of the context for me."* Similarly, five participants found that showing all words was helpful for comprehension. Another issues was the selection of keywords: P27 noted that *"the selected words did not necessarily coincide with [their] interest"* and P42 found the highlighting of words in background conversations confusing.

8.7 Preferred Caption Designs

As an outlook, we asked participants how they would design their own captions. We clustered responses in Table 2. Nineteen participants said they would stick to standard caption with no or almost no modification, largely because this is what they and other viewers are already used to. Twenty-one participants described a design very close to (time-synchronized) keyword highlights, adding some suggestions such as different typesetting. Eighteen participants listed additional elements to be included or changed in the captions, for example, different colors to distinguish speakers or background information on certain words.

9 Discussion

By providing insights into the user perspective on captioned videos, we support researchers and practitioners in motivating users to embed learning activities into their everyday viewing experiences. In particular, the opportunities and challenges we identified—such as the need for context, habits, distractions, and the potential to focus attention—inform the design of captioning for learning, comprehension, and entertainment.

9.1 Distractions Outweigh the Potential of Enhanced Captions for Entertainment and Comprehension

Although *Keyword Highlights* and *Timed Keyword Highlights* performed better than or similar to *Standard Captions* on various measures, the overall ranking in Figure 5 clearly shows that standard captions were the go-to solution in terms of comprehension and entertainment; only in the learning dimension, *Timed Keyword Highlights* overtook *Standard Captions*. Specifically, *Keyword Highlights* and *Timed Keyword Highlights* were similarly attractive alternatives on the pragmatic subscale of the User Experience Questionnaire and were rated higher on the hedonic subscale. Similarly, the number of participants describing their preferred captions as a variant of *Standard Captions* or *(Timed) Keyword Highlights* captions was almost the same. Still, it seems that due to the increased potential for distractions, the two caption variants that used highlighted keywords were not perceived as sufficiently agreeable, innovative, or helpful to overrule the influence of habits and familiarity. The ranking and participant statements further indicate that learners are only willing to accept divided foci of attention in a learning scenario.

Research on visual perception agrees that sudden and easily distinguishable stimuli attract a viewer's attention [6]. Thus, it is unsurprising that a colored and/or suddenly appearing keyword will achieve this. So, while Mirzaei et al. [30] recommended timed keyword captions as a good alternative to standard captions because of the high density of relevant words, our findings suggest that participants did not like the viewing experience with timed changes and bright colors. From a design perspective, less obtrusive highlights, such as bold or italic print, could be used (see also *textual enhancement* strategies [23]).

9.2 Choosing Ideal Keywords is Hard – Optimize Designs for Heuristics and Curricula

We chose our keywords based on a word frequency corpus aligned with estimated language levels. This is a typical approach in language learning and was, for example, also used by Mirzaei et al. [30]. In other projects, keywords were based on expert ratings [14] or a pre-test [34]. However, especially in our interconnected world and

Table 2: Clusters of responses to the question "If you could design your own captions, how would they look?", including exemplary statements

Additional elements – Marking speakers or objects: 7 participants P10: *I would assign colors to the characters so they can be distinguished more easily when several voices overlap*
Additional elements — Translations, explanations, synonyms: 11 participants P19: *They would be very similar to the timed keyword highlights, maybe with a synonym in brackets or including the translation of the word [...]* P21: *With color codes that indicate if the highlighted words are verbs, nouns, etc.* P30: *[...] with other words that are easier to understand and that are synonyms of the [keywords]*
Style suggestions, e.g., fonts: 15 participants P1: *I would focus on the clarity of the subtitles above all* P34: *Simple, either white or yellow with a black contour so the font remains legible on a white background*
(Time-Synchronized) keyword highlights (with minor changes): 21 participants P39: *They would be a combination of standard subtitles with highlighted words in brackets* P45: *maybe putting [keywords] a bit bigger than the other words or with a frame to mark the importance of the word* P54: *They would appear, similar to karaoke, timed to match the pronunciation*
Standard captions (with minor changes): 19 participants P14: *I would simply leave it at the standard because it does a really good job and everyone is used to it* P41: *Traditional captions because not everyone does not know the same words* P48: *The truth is that standard captions are pretty similar to those I would design for my use*

for a ubiquitous language such as English, it is almost impossible to perfectly model a learner's prior knowledge to predict unknown vocabulary. In fact, several participants in our study mentioned that the selected keywords did not match their expectations. Moreover, watching movies is often a social experience including two or more people, and adding another person to the equation complicates the process even further.

This means that keyword highlights will, at most, be an educated guess. But how critical is this, really? We argue that a suitable caption design that balances distractions, context, and focus is more crucial. In particular, we expect that highlighting a few words too many will not have a dramatic impact on the viewing experience, as long as they do not annoy or distract the viewer (as was the case in our study). Consequently, we recommend a conservative selection of keywords. Furthermore, in the movie analyses performed by Andrade [1], a substantial share of the vocabulary was estimated at B2 level or lower, indicating that the number of keywords in most movies will not surpass a certain threshold. To preserve the context, the participants of our study demanded full captions. This is also beneficial with respect to imprecise keyword selection: full captions ensure that false negative keywords (unknown words that are not highlighted) will still be visible, albeit not highlighted.

Alternatively, captioned viewing could be aligned with classroom learning. We suggest a crowdsourced approach to collect target word lists. For example, Culbertson et al. [7] proposed a system for correcting auto-generated captions that could be extended with a feature for learners to highlight words relevant to their language class.

9.3 Limitations and Future Work

Our initial hope was that our caption enhancements would foster learning without causing a negative impact on the viewing experience. If this were the case, there would be no reason for viewers to stick with standard captions. However, enhanced captions were only top-ranked for a learning scenario. This highlights the need for further adaptations to make the viewing experience with enhanced captions similarly enjoyable. Currently, we do not know to what extent this preference was caused by our design choices, such as using the yellow color for highlights. Consequently, future work should analyze the effect of design choices, factoring in findings from label design [21]. We also encountered technical and methodological challenges during the implementation and evaluation of the caption types. Notably, our processing pipeline is not yet fully automated and can, therefore, not be applied at scale. For example, in two of the scenes we used, the lines of two characters partially overlapped. This required swapping some lines for the forced alignment, which our system is currently not capable of doing automatically. In addition, although we aim to support implicit learning in everyday life, we focused on user experience and did not measure learning in detail. A long-term, in-situ study would be necessary to assess learning success with different caption types. Future work should also investigate to what extent the findings hold for languages with larger linguistic differences and different writing systems.

10 Conclusion

In this paper, we implement and evaluate three enhanced caption types that increase the focus on target words in language learning by highlighting and/or displaying words synchronized with the audio track. To gather viewers' opinions on these captions, we conducted an online survey evaluating the user experience, perceived

comprehension, and vocabulary recognition with our enhanced caption types compared to standard captions. We discovered that participants preferred captions with highlights in a learning scenario but felt that they were too distracting for an everyday viewing experience. These findings highlight challenges in the widespread adoption of captions optimized for learning in language learners' everyday lives.

Acknowledgments

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 251654672 – TRR 161.

References

[1] Frank Andrade. 2020. The Best Movies to Learn a Foreign Language According to Data Science. https://towardsdatascience.com/the-best-movies-to-learn-english-according-to-data-science-2dccb4b3ee23

[2] Francesca Bianchi and Tiziana Ciabattoni. 2008. Captions and subtitles in EFL learning: An investigative study in a comprehensive computer environment. EUT-Edizioni Università di Trieste, 69–90.

[3] Yi-Ru Chen, Yeu-Ting Liu, and Andrew Graeme Todd. 2018. Transient but Effective? Captioning and Adolescent EFL Learners' Spoken Vocabulary Acquisition. *English Teaching & Learning* 42, 1 (May 2018), 25–56. https://doi.org/10.1007/s42321-018-0002-8

[4] Anthony A. Ciccone. 2014. Teaching with authentic video: Theory and practice. In *Second language acquisition theory and pedagogy*, Fred R. Eckman, Jean Mileham, Rita Rutkowski Weber, Diane Highland, and Peter W. Lee (Eds.). Routledge, London, 203–216. OCLC: 1100448443.

[5] Myrna Cintrón-Valentín, Lorenzo García-Amaya, and Nick C. Ellis. 2019. Captioning and grammar learning in the L2 Spanish classroom. *The Language Learning Journal* 47, 4 (Aug. 2019), 439–459. https://doi.org/10.1080/09571736.2019.1615978

[6] Maurizio Corbetta and Gordon L. Shulman. 2002. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience* 3, 3 (March 2002), 201–215. https://doi.org/10.1038/nrn755

[7] Gabriel Culbertson, Solace Shen, Erik Andersen, and Malte Jung. 2017. Have your Cake and Eat it Too: Foreign Language Learning with a Crowdsourced Video Captioning System. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 286–296. https://doi.org/10.1145/2998181.2998268

[8] Martine Danan. 1992. Reversed Subtitling and Dual Coding Theory: New Directions for Foreign Language Instruction. *Language Learning* 42, 4 (Dec. 1992), 497–527. https://doi.org/10.1111/j.1467-1770.1992.tb01042.x

[9] Martine Danan. 2004. Captioning and Subtitling: Undervalued Language Learning Strategies. *Meta* 49, 1 (Sept. 2004), 67–77. https://doi.org/10.7202/009021ar

[10] John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham. 2013. Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest* 14, 1 (Jan. 2013), 4–58. https://doi.org/10.1177/1529100612453266

[11] Wendy Fox. 2016. Integrated titles: An improved viewing experience? *Eyetracking and applied linguistics* (2016), 5–30. https://doi.org/10.17169/LANGSCI.B108.233 Publisher: Language Science Press.

[12] Anca Daniela Frumuselu, Sven De Maeyer, Vincent Donche, and María del Mar Gutiérrez Colon Plana. 2015. Television series inside the EFL classroom: Bridging the gap between teaching and learning informal language through subtitles. *Linguistics and Education* 32 (Dec. 2015), 107–117. https://doi.org/10.1016/j.linged.2015.10.001

[13] 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 on - AH2019*. ACM Press, Reims, France, 1–9. https://doi.org/10.1145/3311823.3311865

[14] Helen Gant Guillory. 1998. The effects of keyword captions to authentic French video on learner comprehension. *Calico Journal* (1998), 89–108.

[15] Ching-Kun Hsu, Gwo-Jen Hwang, Yu-Tzu Chang, and Chih-Kai Chang. 2012. Effects of Video Caption Modes on English Listening Comprehension and Vocabulary Acquisition Using Handheld Devices. *Educational Technology & Society* 16, 1 (2012), 403–414.

[16] Johanna W Istanto. 2009. The use of films as an innovative way to enhance language learning and cultural understanding. *Electronic Journal of Foreign Language Teaching* 6, 1 (2009), 278–290.

[17] JASP Team. 2022. JASP (Version 0.16.3)[Computer software]. https://jasp-stats.org/

[18] Geza Kovacs and Robert C. Miller. 2014. Smart subtitles for vocabulary learning. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. ACM Press, Toronto, Ontario, Canada, 853–862. https://doi.org/10.1145/2556288.2557256

[19] Stephen D. Krashen. 2004. *The power of reading: insights from the research* (2nd ed ed.). Libraries Unlimited ; Heinemann, Westport, Conn. : Portsmouth, NH.

[20] Jan-Louis Kruger, Stephen Doherty, Wendy Fox, and Peter de Lissa. 2018. Chapter 12. Multimodal measurement of cognitive load during subtitle processing: Same-language subtitles for foreign-language viewers. In *American Translators Association Scholarly Monograph Series*, Isabel Lacruz and Riitta Jääskeläinen (Eds.). Vol. XVIII. John Benjamins Publishing Company, Amsterdam, 267–294. https://doi.org/10.1075/ata.18.12kru

[21] Ernst Kruijff, Jason Orlosky, Naohiro Kishishita, Christina Trepkowski, and Kiyoshi Kiyokawa. 2019. The Influence of Label Design on Search Performance and Noticeability in Wide Field of View Augmented Reality Displays. *IEEE Transactions on Visualization and Computer Graphics* 25, 9 (Sept. 2019), 2821–2837. https://doi.org/10.1109/TVCG.2018.2854737

[22] Kuno Kurzhals, Emine Cetinkaya, Yongtao Hu, Wenping Wang, and Daniel Weiskopf. 2017. Close to the Action: Eye-Tracking Evaluation of Speaker-Following Subtitles. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM Press, Denver, Colorado, USA, 6559–6568. https://doi.org/10.1145/3025453.3025772

[23] Ryan M LaBrozzi. 2016. The effects of textual enhancement type on L2 form recognition and reading comprehension in Spanish. *Language Teaching Research* 20, 1 (Jan. 2016), 75–91. https://doi.org/10.1177/1362168814561903

[24] Rodrigo Laiola Guimarães, Priscilla Avegliano, and Lucas C. Villa Real. 2016. A Lightweight and Efficient Mechanism for Fixing the Synchronization of Misaligned Subtitle Documents. In *Proceedings of the 2016 ACM Symposium on Document Engineering*. ACM, Vienna Austria, 175–184. https://doi.org/10.1145/2960811.2960812

[25] Sven Lommel, Annouschka Laenen, and Géry d'Ydewalle. 2006. Foreign-grammar acquisition while watching subtitled television programmes. *British Journal of Educational Psychology* 76, 2 (June 2006), 243–258. https://doi.org/10.1348/000709905X38946

[26] Qikun Ma, Shiyang Wang, Jie Liu, and Nianlong Li. 2018. InteractiveSubtitle: Subtitle Interaction for Language Learning. In *Proceedings of the Sixth International Symposium of Chinese CHI on - ChineseCHI '18*. ACM Press, Montreal, QC, Canada, 116–119. https://doi.org/10.1145/3202667.3202685

[27] Paul. L. Markham, Lizette A. Peter, and Teresa J. McCarthy. 2001. The Effects of Native Language vs. Target Language Captions on Foreign Language Students' DVD Video Comprehension. *Foreign Language Annals* 34, 5 (Sept. 2001), 439–445. https://doi.org/10.1111/j.1944-9720.2001.tb02083.x

[28] Rafael Matielo, Raquel Carolina Souza Ferraz D'Ely, and Luciane Baretta. 2015. The effects of interlingual and intralingual subtitles on second language learning/acquisition: a state-of-the-art review. *Trabalhos em Linguística Aplicada* 54, 1 (June 2015), 161–182. https://doi.org/10.1590/0103-18134456147091

[29] R.E. Mayer. 2017. Using multimedia for e-learning. *Journal of Computer Assisted Learning* 33, 5 (Oct. 2017), 403–423. https://doi.org/10.1111/jcal.12197

[30] Maryam Sadat Mirzaei, Yuya Akita, and Tatsuya Kawahara. 2014. Partial and synchronized captioning: A new tool for second language listening development. In *CALL Design: Principles and Practice - Proceedings of the 2014 EUROCALL Conference, Groningen, The Netherlands*. Research-publishing.net, 230–236. https://doi.org/10.14705/rpnet.2014.000223

[31] Holger Mitterer and James M. McQueen. 2009. Foreign Subtitles Help but Native-Language Subtitles Harm Foreign Speech Perception. *PLoS ONE* 4, 11 (Nov. 2009), e7785. https://doi.org/10.1371/journal.pone.0007785

[32] Maribel Montero Perez. 2022. Second or foreign language learning through watching audio-visual input and the role of on-screen text. *Language Teaching* 55, 2 (April 2022), 163–192. https://doi.org/10.1017/S0261444821000501

[33] Maribel Montero Perez, Elke Peters, Geraldine Clarebout, and Piet Desmet. 2014. Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning & Technology* 18, 1 (Feb. 2014), 118–141. http://hdl.handle.net/10125/44357

[34] Maribel Montero Perez, Elke Peters, and Piet Desmet. 2015. Enhancing vocabulary learning through captioned Video: An eye-tracking study. *The Modern Language Journal* 99, 2 (2015), 308–328. https://doi.org/10.1111/modl.12215

[35] Maribel Montero Perez, Wim Van Den Noortgate, and Piet Desmet. 2013. Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System* 41, 3 (Sept. 2013), 720–739. https://doi.org/10.1016/j.system.2013.07.013

[36] Carmen Muñoz, Geòrgia Pujadas, and Anastasiia Pattemore. 2023. Audio-visual input for learning L2 vocabulary and grammatical constructions. *Second Language Research* 39, 1 (Jan. 2023), 13–37. https://doi.org/10.1177/02676583211015797

[37] Nadiia Mykhalevych. 2023. Survey: Why America is obsessed with subtitles. https://preply.com/en/blog/americas-subtitles-use/

[38] Ian Stephen Paul Nation. 2016. *Making and using word lists for language learning and testing*. Vol. 10. John Benjamins Publishing Company Amsterdam.

[39] Paul Nation. 2007. The Four Strands. *Innovation in Language Learning and Teaching* 1, 1 (April 2007), 2–13. https://doi.org/10.2167/illt039.0
[40] Elke Peters. 2019. The Effect of Imagery and On-Screen Text on Foreign Language Vocabulary Learning From Audiovisual Input. *TESOL Quarterly* 53, 4 (Dec. 2019), 1008–1032. https://doi.org/10.1002/tesq.531
[41] Elke Peters and Stuart Webb. 2018. INCIDENTAL VOCABULARY ACQUISITION THROUGH VIEWING L2 TELEVISION AND FACTORS THAT AFFECT LEARNING. *Studies in Second Language Acquisition* 40, 3 (Sept. 2018), 551–577. https://doi.org/10.1017/S0272263117000407
[42] Michael P.H. Rodgers and Stuart Webb. 2020. Incidental vocabulary learning through viewing television. *ITL - International Journal of Applied Linguistics* 171, 2 (Sept. 2020), 191–220. https://doi.org/10.1075/itl.18034.rod
[43] Nathan Sakunkoo and Pattie Sakunkoo. 2009. Gliflix: Using movie subtitles for language learning. In *Proceedings of the 26th Symposium on User Interface Software and Technology. ACM.* https://uist.acm.org/archive/adjunct/2009/pdf/demos/paper171.pdf
[44] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence* 4, 6 (2017), 103. https://doi.org/10.9781/ijimai.2017.09.001
[45] Gregory Taylor. 2005. Perceived Processing Strategies of Students Watching Captioned Video. *Foreign Language Annals* 38, 3 (Oct. 2005), 422–427. https://doi.org/10.1111/j.1944-9720.2005.tb02228.x
[46] (Mark) Feng Teng. 2022. Vocabulary learning through videos: captions, advance-organizer strategy, and their combination. *Computer Assisted Language Learning* 35, 3 (March 2022), 518–550. https://doi.org/10.1080/09588221.2020.1720253
[47] Mark Feng Teng. 2023. Incidental vocabulary learning from captioned videos: Learners' prior vocabulary knowledge and working memory. *Journal of Computer Assisted Learning* 39, 2 (April 2023), 517–531. https://doi.org/10.1111/jcal.12756
[48] Robert Vanderplank. 2016. *Captioned Media in Foreign Language Learning and Teaching: Subtitles for the Deaf and Hard-of-Hearing as Tools for Language Learning.* Palgrave Macmillan UK, London. https://doi.org/10.1057/978-1-137-50045-8
[49] Venkatesh, Thong, and Xu. 2012. Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Quarterly* 36, 1 (2012), 157. https://doi.org/10.2307/41410412
[50] Andi Wang and Ana Pellicer-Sánchez. 2022. Incidental vocabulary learning from bilingual subtitled viewing: An eye-tracking study. *Language Learning* 72, 3 (2022), 765–805.
[51] Yeshuang Zhu, Yuntao Wang, Chun Yu, Shaoyun Shi, Yankai Zhang, Shuang He, Peijun Zhao, Xiaojuan Ma, and Yuanchun Shi. 2017. ViVo: Video-Augmented Dictionary for Vocabulary Learning. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* ACM, Denver Colorado USA, 5568–5579. https://doi.org/10.1145/3025453.3025779

A Survey Measures

Table 3: Questions on subtitles and captions included in the online survey

Question (translated to English)	Question Type
I like to use subtitles/captions very much (any language).	5-point scale
How often have you used subtitles/captions in the past 30 days?	Selection menu
In what situations do you use subtitles/captions? (any language)?	Selection menu + other text field
How do you set subtitles/captions when the video is in a foreign language (any language)?	Selection menu + other text field
If you could design your own subtitles/captions, how would they look?	Text field

B User Study Measures

Table 4: Questions (translated to English) on demographics, English experience, caption habits and preferences, self-assessment of viewing with captions asked in the user study

Measure	Question	Question Type
Demographics	How old are you?	Text Field
Demographics	How do you identify yourself?	Selection menu + other text field
Demographics	In which country do you currently live?	Selection menu + other text field
Demographics	What level of education do you have?	Selection menu + other text field
Demographics	What is your current occupation?	Selection menu + other text field
Demographics	What is your native language?	Selection menu + other text field
English Experience	How often do you speak English?	Selection menu
English Experience	How often do you need to understand English (for example, when reading or on the Internet)?	Selection menu
English Experience	What is your English language level?	Selection menu
Vocabulary Pre-Test	What synonym or definition can you use to meaningfully replace the words in angle brackets in the following sentences?	4 Options per question
Caption Habits & Preferences	I like to use subtitles very much (no matter in which language).	7-point scale
Caption Habits & Preferences	How often have you used subtitles (in any language) in the last 30 days?	Selection menu
Caption Habits & Preferences	How do you set the subtitles if the video is in a foreign language (any language)?	Multiple Choice Selection menu + other text field
Self-Assessment	I understood the language very well.	6-point scale
Self-Assessment	I understood the plot very well.	6-point scale
Self-Assessment	I have the impression that I can learn new words very well with this subtitle variant.	6-point scale

Table 5: Questions (translated to English) on user experience, additional feedback, preferred caption designs, and vocabulary retention asked in the user study

Measure	Question	Question Type
User Experience	Watching the video with this kind of subtitles was very pleasant.	6-point scale
User Experience	UEQ-S [44]	7-Point Likert Scale
User Experience	I can very well imagine using this kind of subtitles myself.	6-point scale
User Experience	I really like this subtitle variant overall.	7-point scale
Additional Feedback	Is there anything else you would like to say?	Text field
Self-Assessment	How much did you pay attention to the following aspects while watching the videos? (Scene understanding, Learning new words, Entertainment)	6-point scale for each
User Experience	Please sort all subtitle variants according to how well you like them if the focus is on learning new vocabulary.	Option to sort all 4 variants
Additional Feedback	Why did you sort the variants in this way?	Text field
User Experience	Please sort all subtitle variants according to how well you like them if the focus is on entertainment/pleasure.	Option to sort all 4 variants
Additional Feedback	Why did you sort the variants in this way?	Text field
User Experience	Please sort all subtitle variants according to how well you like them when the focus is on scene comprehension.	Option to sort all 4 variants
Additional Feedback	Why did you sort the variants in this way?	Text field
Desired Captions	If you could design your own subtitles, what would they look like?	Text field
Vocabulary Retention	What synonym or definition can you use to meaningfully replace the words in angle brackets in the following sentences?	4 Options per question
Additional Feedback	Is there anything else you would like to say?	Text field

C Caption Processing Pipeline

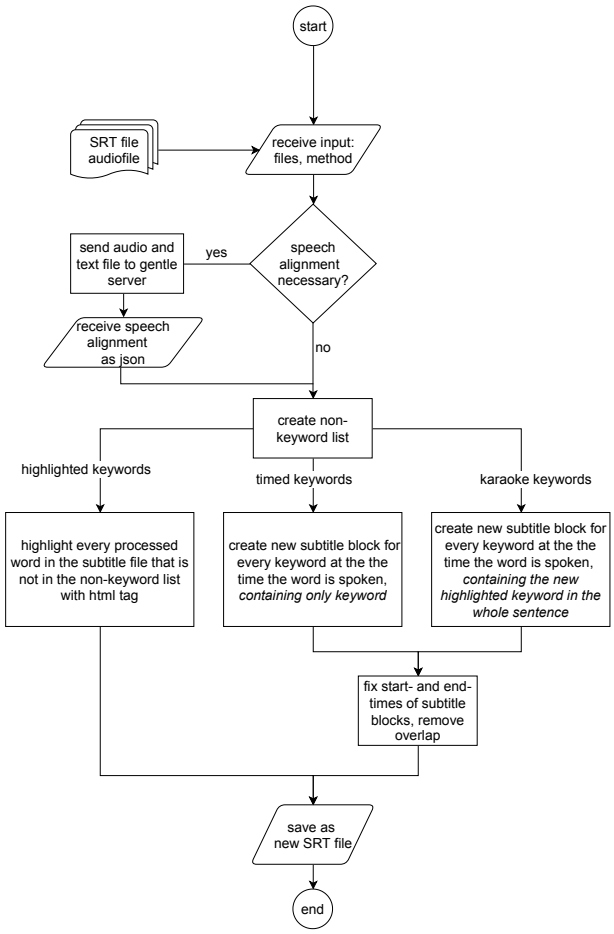

Figure 6: Processing pipeline for subtitle files.