Sentence complexity as an indicator of L2 learner's listening difficulty

Maryam Sadat Mirzaei[1] and Kourosh Meshgi[2]

Abstract. This paper investigates the effect of sentence complexity, specifically lexical and syntactic surprisal, on L2 listening difficulty. Psycholinguistic studies revealed that surprisal cases correlate with textual comprehension difficulty. Based on surprisal theory, these cases are less probable or expected, considering the precedent context, thus require more complex processing to comprehend. Little is known about the influence of the surprisal factor on L2 listening comprehension. We aim to examine this effect and propose to include these cases in captioning to assist L2 listeners. Since conventional captions include the whole transcript, we use Partial and Synchronized Caption (PSC) with limited textual clues, which allows for highlighting surprisal cases to reduce ambiguity. In our experiment, intermediate learners of English (undergraduates) were asked to transcribe and paraphrase videos containing surprisal cases. Results revealed that learners faced difficulty when encountering surprisal, which was partially addressed with the help of PSC, yet more assistance was required.

Keywords: partial and synchronized caption, surprisal model, parsing complexity.

1. Introduction

Investigating appropriate methods to teach L2 listening is a continuing concern given that listening has been long considered as a passive skill (Osada, 2004). Several factors are known to make L2 listening difficult, including acoustic, lexical, syntactic, and content-related features (Bloomfield et al., 2010). Previous research has investigated the influence of syntactic features on reading difficulty, but this aspect is not adequately considered in L2 listening. One of the elements involved

1. RIKEN AIP, Tokyo, Japan; maryam.mirzaei@riken.jp; https://orcid.org/0000-0002-0715-1624
2. RIKEN AIP, Tokyo, Japan; kourosh.meshgi@riken.jp; https://orcid.org/0000-0001-7734-6104

How to cite: Mirzaei, M. S., & Meshgi, K. (2020). Sentence complexity as an indicator of L2 learner's listening difficulty. In K.-M. Frederiksen, S. Larsen, L. Bradley & S. Thouësny (Eds), *CALL for widening participation: short papers from EUROCALL 2020* (pp. 227-232). Research-publishing.net. https://doi.org/10.14705/rpnet.2020.48.1193

in sentence complexity is surprisal, which relates to the predictability of a word in the context, with a highly probable word being easier to process. According to the expectation-based model for syntactic comprehension, one measures the probability of the next input based on the preceding context (Levy, 2008). Studies using fMRI, EEG, and eye-tracking provide evidence for the effect of surprisal on working memory load, reading time, and comprehension (Smith & Levy, 2013). However, little is known about how this factor affects L2 listening.

In this study, we investigate whether syntactic and lexical surprisal affects L2 listening difficulty and propose the inclusion of this factor in PSC to facilitate listening. PSC is a captioning tool that automatically detects difficult words/phrases, includes them in the caption, and removes trivial cases (partial). It also synchronizes each word with the relevant speech segment (word-level synchronization). PSC aims to decrease dependence on the caption and promote listening over reading (Mirzaei, Meshgi, & Kawahara, 2018). Only acoustic (speech rate, breached boundaries, acoustically similar words) and lexical factors (word frequency, specificity) are used in PSC, yet sentence complexities are not addressed. In this study, we focus on syntactic surprisal using the structural confusion of a sentence, discovered by a probabilistic grammar/parser. We also measure lexical surprisal, utilizing the probability of the next word based on a corpus-based N-gram. The words with high surprisal scores are selected to be included in PSC.

Figure 1. Screenshot of a surprisal case appearing in PSC

Original sentence: "*Scientists were unprepared for this tsunami of doubt and questions and distrust.*"
Heidi Larson | TEDMED2020

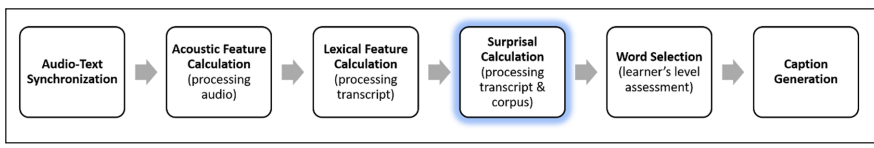
Figure 1 above shows a TED talk with PSC that includes a surprisal case ([tsunami of] doubt). The selective nature of PSC allows for highlighting challenging aspects of listening. By adding surprisal, we aim to facilitate recognition and comprehension, decrease cognitive load, and foster ambiguity resolution.

2. Lexical and syntactic surprisal for L2 listening

There is a strong correlation between lexical or syntactic surprisal with the required effort for parsing and processing the sentences (e.g. "the horse raced past the barn fell"). This notion is based on surprisal theory (Levy, 2008), which assumes that a word's predictability can determine difficulty. In this view, the cognitive effort it takes for the learner to process a word is proportional to its surprisal (Hale, 2001).

Speech is transient, and we can assume that when a learner encounters a word that is different from what she/he expects to hear, the attention is confined, leading to confusion, cognitive overload, and misrecognition. A similar situation can happen when a learner tries to match a preferred sentence structure to an input speech and finds a mismatch. To investigate this hypothesis, we included the words having a high surprisal score to the PSC generated for TED talks to be used as material for L2 listeners (Figure 2).

Figure 2. The procedure for generating PSC including surprisal cases



We used N-gram surprisal and PCFG[3] surprisal to detect lexical and structural surprisal cases. N-grams are calculated on TED corpus using *KenLM*, and the lexical surprisal is calculated as the negative log probability of the word given the previous *N-1* words. A *PCFG-based incremental parser* (van Schijndel, Exley, & Schuler, 2013) is employed to determine the dependency relations of previous words. Similar to how humans comprehend the input, an incremental parser integrates incoming words in a syntax that fits the preceding context. Surprise arises when the input word changes the probability distribution over the possible parses, namely the expectation of the parser about the underlying syntax. Each word of an

3. Probabilistic Context-Free Grammar

N-gram ending with a lexical surprisal is included in PSC to facilitate recognition and avoid surprises. For syntactic surprisal, a part of the sentence whose parse tree is changed drastically when processing the latest input is considered surprising and is shown in the caption all at once.

3. Preliminary evaluation and discussion

Our participants were 17 Japanese and nine Chinese engineering undergraduates who were intermediate learners of English (aged 19-21 with 520~725 TOEIC[4] scores ≈ CEFR[5] B1). We selected 20 clips from TED videos, on average 37 seconds, each including one surprisal case (ten lexical and ten syntactic cases) and one easy case (control). The easy cases were selected using PSC's difficulty measures (i.e. words that are automatically omitted for being trivial). When selecting the surprisal cases, we made sure that acoustic-related difficulties are not present.

The participants were asked to watch each video segment and fill the blanks for the 2~3 words in the last sentence, when the video was paused. Subsequently, they were asked to paraphrase that sentence. The purpose was to check how accurately the learners can transcribe/paraphrase easy versus surprisal cases.

Figure 3 shows participants' correct and incorrect answers on easy versus surprisal cases as well as their scores of the paraphrasing task. As the figure suggests, participants could transcribe the easy cases significantly better than the surprisal ones. Data shows that learners faced slightly more difficulties in transcribing lexical surprisals as compared with syntactic surprisals. However, this difference was not statistically significant (p=0.35). In the paraphrasing task, learners' scores indicate that more difficulty is associated with syntactic surprisal case. It can be argued that lexical surprisal leads to more misrecognition (e.g. [tsunami of] 'doubt', transcribed as 'that'), while syntactic surprisal makes comprehension more difficult.

Finally, to check if the inclusion of surprisal cases in PSC can assist learners with listening, we asked the participants to re-paraphrase the target segment after watching it with enhanced PSC (showing surprisal cases). The results are demonstrated in Figure 4, which indicates that PSC could significantly facilitate comprehension for lexical surprisal cases (p<0.05). Although including syntactic cases in PSC resulted in better scores, the improvement was not significant (p=0.06). Moreover, participants' overall scores reflect that a better sort of scaffold

4. Test of English for International Communication
5. Common European Framework of Reference for languages

is necessary to help them improve their performance. This finding suggests that merely showing these cases in PSC was not adequate for alleviating comprehension difficulty. Thus a better method should be considered to help learners comprehend structural surprisals. Generating shorter or simplified sentences and presenting them along with the original one in PSC could be one way to address this issue. Furthermore, repeating the experiment with control and treatment groups and learners of different proficiency levels can provide insights to design better tools.

Figure 3. Participants' scores on transcription and paraphrasing of easy versus surprisal cases

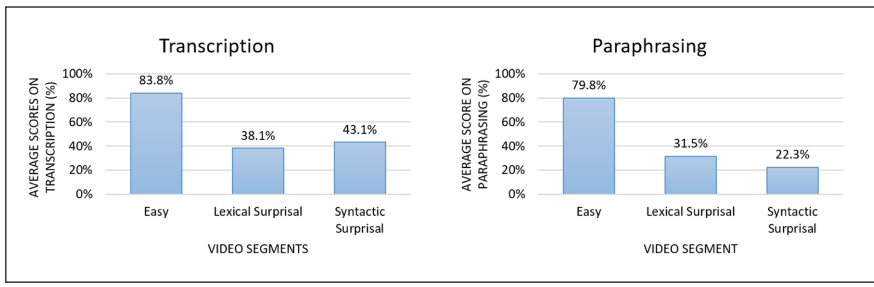
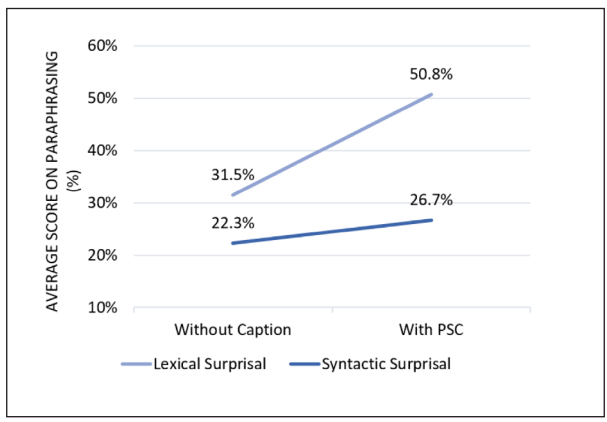


Figure 4. The score of participants on paraphrasing task with/without using PSC



4. Conclusions

We investigated the influence of syntactic and lexical surprisal on L2 learners' listening and found that the existence of surprisal cases leads to difficulty in

recognition (cloze-test transcription) and comprehension (paraphrasing test) of the speech input. Findings revealed that the inclusion of these cases into PSC is more helpful with lexical surprisal cases than structural ones. However, further evaluation is necessary to find in what ways, including these cases into PSC, can foster listening. Additionally, more conclusive results could be gained using eye-trackers to investigate the learner's fixation and cognitive load when surprisal cases are presented in the caption. Future work should consider a more effective method to address these cases. Simplification of the syntactic surprisal cases and adding them to caption could be one approach to consider.

References

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). *What makes listening difficult? Factors affecting second language listening comprehension*. Maryland University College Park. https://doi.org/10.21236/ada550176

Hale, J. (2001). A probabilistic early parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1-8). https://doi.org/10.3115/1073336.1073357

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177. https://doi.org/10.1016/j.cognition.2007.05.006

Mirzaei, M. S., Meshgi, K., & Kawahara, T. (2018). Exploiting automatic speech recognition errors to enhance partial and synchronized caption for facilitating second language listening. *Computer Speech & Language*, *49*, 17-36. https://doi.org/10.1016/j.csl.2017.11.001

Osada, N. (2004). Listening comprehension research: a brief review of the past thirty years. *Dialogue*, *3*(1), 53-66.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302-319. https://doi.org/10.1016/j.cognition.2013.02.013

Van Schijndel, M., Exley, A., & Schuler, W. (2013). A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, *5*(3), 522-540.

Published by Research-publishing.net, a not-for-profit association
Contact: info@research-publishing.net

CALL for widening participation: short papers from EUROCALL 2020
Edited by Karen-Margrete Frederiksen, Sanne Larsen, Linda Bradley, and Sylvie Thouësny

Publication date: 2020/12/14

Typeset by Research-publishing.net
Cover theme by © 2020 Marie Flensborg (frw831@hum.ku.dk), based on illustration from freepik.com
Cover layout by © 2020 Raphaël Savina (raphael@savina.net)

ISBN13: 978-2-490057-81-8 (Ebook, PDF, colour)

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2020.