卒業論文

リアルタイム翻訳字幕生成における翻訳タイミングの評価

指導教官　　村上 陽平 教授

立命館大学　情報理工学部
先端社会デザインコース　4回生
2600190086-1

笠井 直晴

2023年度（秋学期）卒業研究3（CH）
令和6年1月31日

リアルタイム翻訳字幕生成における翻訳タイミングの評価

笠井直晴

内容梗概

コロナウイルスにより，授業や仕事などが遠隔で行なわれる機会が増えている．それに伴い，外国人と遠隔会議システムを通して話す機会が多くなってきているが，英語の運用能力の低い人達は外国人の発話内容を理解できず，たとえ運用能力が高かったとしても，ネイティブの英語には慣れておらず聞き取れないという事がある．この問題を解決するためにニューラルネットワークをベースとした音声認識と機械翻訳を用いたリアルタイム翻訳ツールが多数公開されている．

しかしながら，既存のリアルタイム翻訳では，話者が発話を終えるまで音声認識を継続するため，発話完了してから翻訳文が提示されるまでタイムラグが生じるので，画面と翻訳文が同期せず，聞き手にとって読みにくかったり認知負荷が大きかったりする．

そこで， 本研究では聞き手の認知負荷を抑えるために，リアルタイム翻訳の翻訳タイミングとその原文の編集方法の評価を行う．

具体的には，既存のリアルタイム翻訳と同様に，一文の発話完了時の翻訳のタイミングを変えて被験者の認知負荷の測定を行う．本手法の実現にあたり，取り組むべき課題は以下の 2 点である．

発話の分割単位

翻訳タイミングを決定するために，単語数や秒数，ピリオドなど，発話を区切る最適な発話の分割単位を見つける必要がある．発話の分割単位が短いほど，発話と画面が同期し理解しやすくなるものの，画面上の翻訳文が不完全な状態で提示されるため認知負荷が高くなると考えられる．一方，単位が長くなるほど，タイムラグが生じ，理解しにくくなるものの，翻訳文が完全に近い状態で提示され認知負荷が低くなると考えられる．

字幕生成方法

翻訳字幕の生成方法には，認識した発話単位を逐次翻訳し生成する方法と，認識した発話単位を連結しながら翻訳する方法がある．前者は一度翻訳した部分は固定のため認知負荷が低いものの，翻訳文が不完全で理解しづらく，後者は翻訳文が徐々に完全な状態に近づくため理解しやすいものの，一度翻訳した部分も変化するため認知負荷が高くなると考えられる．理解しやすく

認知負荷も低い字幕生成方法と発話の分割単位の組み合わせを見つける必要がある.

一つ目の音声の発話の分割単位に対して，一文，10 単語，文節の 3 種類の単位を比較して検証を行う．一文は完全な文を翻訳できるが，画面とのタイムラグが生じる．10 単語は画面に同期しやすいが，不完全な文の翻訳となる．リアルタイム性と翻訳の正しさを両立するために，意味的な塊として文節も発話の分割単位として評価する.

二つ目の課題に対して，認識した発話単位をその都度翻訳する逐次翻訳方式と，認識した発話単位を文末がくるまで連結して翻訳する連結翻訳方式の二種の翻訳方式を用いる．この二種の翻訳方式と三種の発話の分割単位を組み合わせて，計六つの組み合わせで字幕を生成する．なお，発話の分割単位が一文の場合，逐次翻訳方式と連結翻訳方式は同じになるため，実際には五つの字幕生成手法を比較評価する．評価では，各手法を適用した動画を視聴後，被験者の理解度と認知負荷の測定を行う.理解度は,動画の内容に基づく問題の正解率で測定し，認知負荷は NASA－TLX を用いて測定し，最適な翻訳タイミングの評価を行った.本研究の貢献は以下の通りである.

発話の分割単位

動画を視聴後理解度を確認する問題と認知負荷を確認するアンケートを行った所,スライド重視の動画が字幕重視の動画より平均正答率が 40%向上し，p 値が 0.01 で有意な差があることが分かった．内容理解度においては発話の分割単位での手法別で有意な差なしと確認出来た．また，認知負荷でも同様に有意な差なしと確認出来た.

字幕生成方法

動画を視聴にあたり被験者がどのぐらい認知負荷がかかっていたのか測定を行うアンケートの行った結果逐次翻訳が連結翻訳よりも，認知負荷の平均の重みが 32%低い事が確認され，p 値は 0.007 で有意な差ありと確認出来た．また，内容理解度における逐次翻訳と連結翻訳では有意な差なしと確認出来た．このことから，逐次翻訳の提案手法が認知負荷においては有効であることが明らかになった.

Evaluation of Timing for Real-Time Translated Subtitles

Tadaharu Kasai

Abstract

Coronavirus, there are more and more opportunities for classes and work to be conducted remotely. People with limited English skills cannot understand what foreigners say, and even if they have good English skills, they may not be able to understand native English speakers because they are not accustomed to it. To solve this problem, many real-time translation tools based on neural network-based speech recognition and opportunity translation have been released. However, existing real-time translation tools continue speech recognition until the speaker finishes speaking, which causes a time lag between the completion of speech and the presentation of the translated text, making it difficult for the listener to read and imposing a heavy cognitive load. Therefore, in order to reduce the cognitive load on listeners, this study evaluated the timing of real-time translation source text. Specifically, as with existing real-time translation, the timing is shifted at sentence completion to measure the subject's cognitive load. In implementing this method, the following two issues needed be addressed.

Speech segmentation unit

In order to determine the timing of translation, it is necessary to find the optimal recognition unit for speech, such as the number of words, seconds, or periods that delimit speech. The shorter the recognition unit, the easier it is to understand, but the cognitive load is considered higher because the translated text on the screen is presented in an incomplete state. On the other hand, the longer the recognition unit, the more difficult it is to understand, but since the translated text is presented in a nearly complete state, the cognitive load is considered low.

Subtitle generation method

There are two methods of generating translated subtitles sequential translation of recognized speech and translation by concatenating recognized speech. First method, the translated part is fixed and the cognitive load is low, but the translated sentence is incomplete and difficult to understand. Second method, the translated sentence gradually approaches a complete state and is easy to understand, but the cognitive load is high because the translated part changes. It

is necessary to find a combination of subtitle generation method and audio recognition unit that is easy to understand and has low cognitive load. The first issue of speech recognition was tested by comparing three types of units: one sentence, 10 words, and phrases. The one-sentence unit could translate the sentence into a complete sentence, but there was a time lag with the movie; the 10-word unit can be easily synchronized with the movie. To achieve both real-time performance and Real-time translation accuracy, we also evaluate the recognition as a semantic. For the second issue, we used two types of translation methods: a sequential translation method that translates each recognized speech unit, and a concatenated translation method that translates each recognized speech unit until the end of the sentence. These two translation methods and three types of speech recognition were combined to generate subtitles using a total of six combinations. In the evaluation, we measure the subject's comprehension and cognitive load after viewing a video to which each method was applied. Comprehension was measured by the percentage of correct answers to questions based on video content, and cognitive load was measured using the NASA-TLX. The contributions of this study are as follows

Speech segmentation unit

After viewing the videos, we conducted a comprehension test and a cognitive load test, and found that the average percentage of correct responses was 40% higher for the slide-oriented videos than for the subtitle-oriented videos, with a significant difference at a p-value of 0.01. In terms of content comprehension, there was no significant difference between the two methods in terms of the segmentation of speech. The same was true for the cognitive load.

Subtitle generation method

The results of a questionnaire survey measuring the degree of cognitive load placed on the subjects as they watched the video showed that the mean weight of cognitive load was 32% lower for the sequential translation than for the consolidated translation, with a p-value of 0.007, indicating a significant difference. There was no significant difference translation in terms of content comprehension. This indicates that the proposed method for sequential translation is effective in terms of cognitive load.

リアルタイム翻訳字幕生成における翻訳タイミングの評価

目次

第1章 はじめに

近年，仕事の会議や授業を遠隔で行う機会が増えている．それに伴い外国人と遠隔会議システムを通して仕事を行ったり，英語の講義を受けたりする事が容易になっている．しかしながら，そのような機会が増えても英語の運用能力が低い人は外国人の発話内容を理解できない．また，たとえ英語の運用能力が高くとも，ネイティブの英語を聞き慣れていなければ，外国人の発話内容が上手く聞き取れない事がある．このような問題を解決するために，Google 翻訳や MicrosoftTranslator などのリアルタイム翻訳が有効である．しかしながら，既存の翻訳ツールは一文ずつ翻訳を行っているので，タイムラグが生じるので画面と翻訳文が同期できず聞き手側にとって読みにくく，認知負荷が高いものとなっている．

そこで，本研究では，聞き手にとって認知負荷の低い区切り方と翻訳結果の見せ方を明らかにすることを目的とする．具体的には，字幕のリアルタイム性を高めるために，一文の認識を待たずに，文を分割して順次翻訳を行う．分割の単位を単語数，カンマ，ピリオドの 3 種に分け，さらに，分割したフレーズの翻訳手法も逐次翻訳と，翻訳済みのフレーズも連結して翻訳する連結翻訳に分けて認知負荷を比較する．本研究で取り組むべき課題は以下の 2 点である．

発話の分割単位

翻訳タイミングを決定するために，単語数や秒数，ピリオドなど，発話を区切る最適な発話の分割単位を見つける必要がある．分割単位が短いほど，発話と画面が同期し理解しやすくなるものの，画面上の翻訳文が不完全な状態で提示されるため認知負荷が高くなると考えられる．一方，単位が長くなるほど，タイムラグが生じ，理解しにくくなるものの，翻訳文が完全に近い状態で提示され認知負荷が低くなると考えられる．

翻訳字幕の生成方法

翻訳字幕の生成方法には，認識した発話単位を逐次翻訳し生成する方法と，認識した発話単位を連結しながら翻訳する方法がある．前者は一度翻訳した部分は固定のため認知負荷が低いものの，翻訳文が不完全で理解しづらく，後者は翻訳文が徐々に完全な状態に近づくため理解しやすいものの，一度翻訳した部分も変化するため認知負荷が高くなると考えられる．理解しやすく認知負荷も低い字幕生成方法と発話の分割単位の組み合わせを見つける必

要がある.

以下，本論文では，まず初めにリアルタイム翻訳の字幕生成のタイミングとは何かの関連研究と共に説明し，動画の字幕生成の手法を説明したのち，検証実験の説明とその結果最後に考察を行う.

第2章 関連研究

2.1 同時通訳の訳出タイミング

この研究では，音声翻訳技術は向上しているが，長い発話の場合，一文が比較的に長いため翻訳開始までの時間が長くなるという課題が存在している．また同時通訳データから訳出タイミングを得るためには話者と通訳者の発話を 0.5 秒以上の無音区間で分割した通訳者の発話を発話単位と決める．その発話単位をと話者が発した意味と同じ発話を人手で対応付けを行う．対応づけた後訳出タイミングを定める．同時通訳者は意味のまとまった段階で通訳を開始することを考慮している[1]．

Sridhar らも英西タスクにおいて，いくつかの訳出タイミング法の比較を行い，その結果として接続詞と句読点を利用した訳出タイミング法の性能が高かったと報告している[2]．

この研究は，リアルタイムで翻訳する自動音声翻訳システムに焦点を当てている．従来の研究では，同時通訳者がリアルタイムで理解しやすくするためさまざま工夫が行われていることを知られているが，これを考慮した機械翻訳システムの学習は十分に行われていない．そのためこの研究では，同時通訳者による同時通訳データを学習プロセスに取り入れている．その結果，システムの翻訳スタイルが経験豊富な同時通訳者に近づくことが示されている．自動評価指標によると，本システムは 1 年の同時通訳者の経験に匹敵する性能を達成している[3]．

この論文では，グローバリゼーションの進展により異なる言語や文化とのコミュニケーションの必要性が増している中，機械による同時通訳システムが将来必要であると述べられている．そのため，高コストの人間による翻訳に対する代替手段として，統計的認識と翻訳技術を基にした同時通訳システムが提案されている．技術や改善点，ユーザーフレンドリーな結果の提供に関する考察が行われている．評価結果から，機械が既に理解可能な同時通訳を実現できることが示唆されている．機械のパフォーマンス向上の余地がある一方で，人間の実行には認知上の制約があると結論付けされている[4]．

先行研究では発話しなくなったタイミングで区切り翻訳を行っているため，発話をしなくなるまでが長い時は翻訳結果の文が長く結果的に読み手側にとっ

て負荷が大きいものとなってしまう．また通訳目的で研究しているため，発話者と通訳者の発話が混ざり聞き取りにくい物となってしまう．そのため，本研究では認知負荷を低くする発話のタイミングとオンラインに向けて字幕の表示方法の分析を行いたい．

第3章 リアルタイム翻訳字幕生成

本章では，リアルタイム翻訳の字幕生成プロセスについて説明する．生成プロセスは発話文の分割，分割した発話の翻訳，翻訳結果の字幕表示の 3 つのステップに分けられる．

3.1 リアルタイム翻訳

リアルタイム翻訳とは簡単に説明を行う．図 1 の様にまず，話者が発話した内容を認識する．その後認識した発話を発話単位で区切る．その後発話単位で区切ったものをどの様にして翻訳機にかけるか検討する．その後，翻訳を行い翻訳結果を字幕として表示を行う．本研究では，図 1 で青色で囲っている部分は既存手法を用いている．黒色で囲っている音声の発話単位で区切る，区切ったものをどうすれば，読み手側にとって認知負荷が少ない区切り方になるのか，又はより内容を理解できる区切り方ができるのか最適な手法を考える．

図 1　リアルタイム翻訳の流れ

3.2 文分割

翻訳精度を高めるには，一文の発話が完了してから翻訳するのが最適であるが，発話の完了まで翻訳を待つと字幕のリアルタイム性が損なわれる．字幕のリアルタイム性を高めるには，発話文を分割し順次翻訳して表示しなければならない．長く分割するほど翻訳精度は高まるが，リアルタイム性を高めるために，発話文の最適な分割の長さを明らかにする必要がある．本研究では，分割の単位として，意味的な区切りであるカンマ，ピリオドに加えて，文の構造や意味に依存しない特定の単語数を分割単位とする．

3.3 発話翻訳

発話翻訳手法として逐次翻訳と連結翻訳の 2 つを用いる．逐次翻訳とは，発話単位で区切ったものを認識するとその都度翻訳を行い，その結果を字幕に表

示するものである．一方，連結翻訳とは発話単位で区切ったものをピリオドを認識するまで区切ったものを連結していき，その都度翻訳を行い，字幕に表示するものである．逐次翻訳と連結翻訳がどの様に認識し表示を行うか簡単に説明を行う．カンマ区切りの逐次翻訳とカンマ区切り連結翻訳で示したものが図 2，図 3 となる．

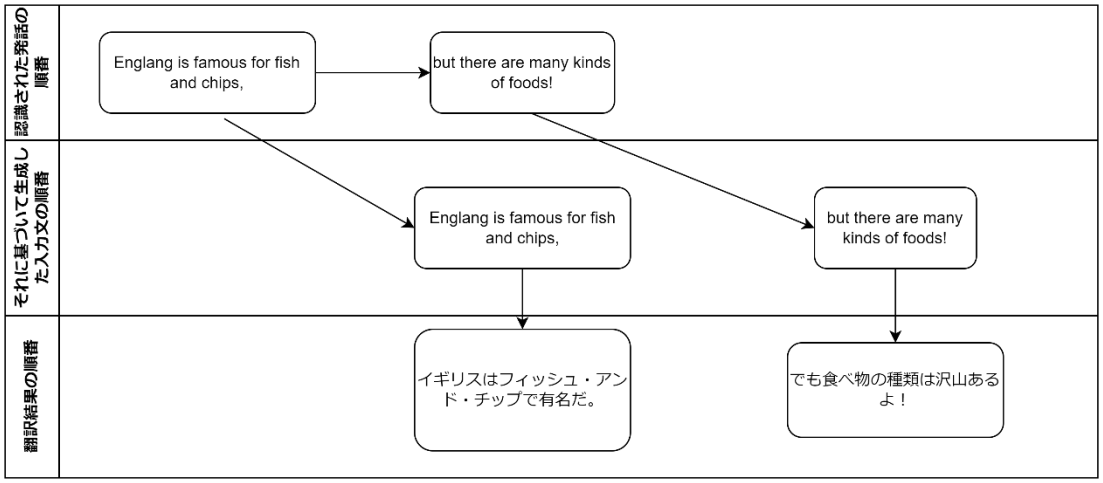

図 2　逐次翻訳の概要

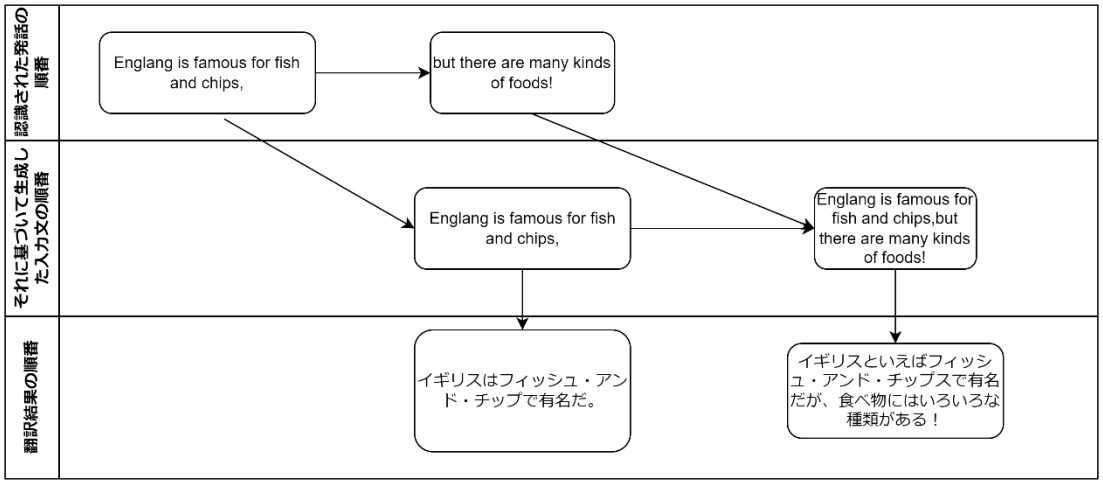

図 3　連結翻訳の概要

3.4 字幕表示

字幕の表示については，既存研究により読みやすいデザインが明らかになっている．表 1 に読みやすい字幕の基準である一行当たりの文字数を示す． 表 1 に記載されている数値は実験を行った 19 人の被験者が動画を視聴しその際に見やすかった字幕を選択し,その字幕のデザインを選んだ人数の結果となっている．表 1 から 1〜2 行数で一行あたりの文字数は 10 文字，15 文字が見やすい字幕となっているため，基本的に字幕の長さはこの範囲に収まるようにした．ただし，ピリオドを分割単位としたり，連結翻訳を行う場合，字幕が必然的に長くなるため，この場合は読み手にとって読みやすい字幕の長さに調節している[5].

表 1　読みやすい字幕のデザイン

行数	一行あたりの文字数				
	10	15	20	10~15	15~20
1	1	1	2	-	-
2	3	3	-	1	1
3	-	1	2	-	-
1~3	-	-	1	-	-
2~3	2	1	-	-	-

第4章 比較実験

本章ではどの発話の分割単位と発話単位のペアが認知負荷が少なく動画の内容をより理解できるのか検証するために行った実験の説明をする．

4.1 参加者

実験に協力していただいたのは，字幕に集中して動画を見て貰うため英語の運用能力は気にせず日本人のみで 12 名である．

4.2 動画

認識した発話単位をその都度翻訳する逐次翻訳方式と，認識した発話単位を文末がくるまで連結して翻訳する連結翻訳方式の二種の翻訳方式を用いる．この二種の翻訳方式と三種の発話の分割単位を組み合わせて，計 5 つの組み合わせで内容が違う 2 つの英語の動画字幕生成と既に翻訳したものを一番見やすい形で字幕にされている YouTube の動画計 12 の字幕動画を生成した．
2 つの英語の動画についてだが，1 つ目の動画が基本的に話者が話しているだけの動画となっており，スライドがあまりない動画となっている．今後動画(スライドなし)と記載する．2 つ目の動画は話者が話しているのにスライドがメインとなっている動画となっている．今後動画(スライドあり)と記載する．1 つの動画に対しての手法を分かりやすく纏めたものが表 2 となっている．

4.3 字幕

字幕についてだが，表 2 に示すように，3 種の分割単位と 2 種の発話翻訳手法の組み合わせでユーザの認知負荷を測定し，最適なリアルタイム翻訳字幕の翻訳タイミングを明らかにする．実験を行う上で分かりやすくするために A～E で振り分けている．また，字幕は 3.4 で説明している様に 1~2 行数で 10~15 単語数で表示していた．分割単位がカンマ，ピリオドの場合は翻訳を行う単語数が多くなるため翻訳結果での単語数も多くなるのでその時はなるべく見やすい様に調整を行った．字幕に見せ方はリアルタイムなので話者が話した会話を認識後翻訳を行っているので，話者が話している時の字幕はその前に認識した発話を翻訳した結果が字幕として表示されている．

表 2　リアルタイム翻訳字幕の生成手法の組み合わせ

分割単位 発話翻訳	10 単語数	カンマ	ピリオド
逐次翻訳	B	D	A
連結翻訳	C	E	

4.4 内容理解度チェック

理解度を測定するために問題を出題している．問題は公平性を保つために採点者や書き方によって点数が曖昧になってしまう記述問題は採用しておらず，選択式のみの問題となっている．単一選択問題では内容を理解していなくても，運で正解してしまう恐れがあったため，全ての問題は 5 択の正解数と回答数を 2 で固定をした問題となっている．

4.5 認知負荷の測定

認知負荷を測定するために動画を視聴後 NASA－TLX を用いている．

4.5.1 NASA-TLX

NASA-TLX（NASA Task Load Index）は，認知負荷を評価するためのツールの一つである．NASA-TLX は，複数の認知的な負荷要因を評価し，タスク全体の主観的な負荷を定量化するために使用される．具体的には，**エラー! 参照元が**

表 3 NASA-TLX 尺度の定義

精神的な要求	動画は見やすかったか見にくかったか
身体的な要求	動画を見ている時とアンケートの最中どれだけ身体を動かしましたか？
時間的な要求	ゆっくりとくつろぎながら字幕を見れたか，又は必死に字幕を見ていたか
努力	問題を解くにあたりあなたはどの程度一生懸命に行いましたか
効率	問題を解くにあたりどの程度正解していると思いますか
不満	動画を見ている間不安やがっかり，いらいら，ストレス感，あるいは逆に安心感や満足感，リラックスをどの程度感じましたか

見つかりません。に示すように， 6 つの要因を評価する[6].

これらの要因に対する被験者の評価は，各要因ごとに 0 から 100 で評定点を得て，その平均値を算出するものである．平均値は一対比較を用いた重み付け係数による平均値 WWL を用いるのが基本的な方法である．この方法で，六つの尺度の全てのペア 15 組について,より負荷が大きいと思う方を選んでもらい各尺度の選択した回数（0~5）をその尺度の重み係数ω_iとする．各尺度の評定値ν_iとすると，WWL は下記の数式で算出され，1~100 の値となる．分母の重み係数ω_iの総和はペアの数と同じ 15 である．

$$WWL = \frac{\sum_{i=1}^{6}(\omega_i \times \nu_i)}{\sum_{i=1}^{6}\omega_i}$$

この評定値を本研究における認知負荷の少なさに用いる．

NASA - TLX の各要因の評定は 0～100 を 20 段階評価したものであるが，被験者が評価しにくいため，本研究では 0 ～ 5 の 6 段階評価を採用した．なお，20 段階評価に換算するために， 0 は 5 ，1 は 20， 2 は 40，3 は 60， 4 は 80， 5 は 100 と重み付け係数の計算を行った．

その後 15 ペアを表示し，各尺度でより負荷の高い尺度を選択してもらった．被験者は，内容理解度の問題と NASA-TLX の質問を GoogleForms にて回答する．また本研究は動画を視聴しているだけなので身体的な要求の尺度に関しては測定が難しいため，結果の際に重要視していない．

4.6 実験計画

認識した発話単位をその都度翻訳する逐次翻訳方式と，認識した発話単位を文末がくるまで連結して翻訳する連結翻訳方式の二種の翻訳方式を用いる．この二種の翻訳方式と三種の発話の分割単位を組み合わせて，計六つの組み合わせで内容が違う英語の動画(スライドなし)と(スライドなし)の 2 つ分の字幕を生成し用意した．

被験者には字幕を生成する際に手法が異なる(スライドなし)と(スライドあり)の 2 つの英語の動画を視聴して貰う．その際，動画は字幕に集中して見て貰う，動画の理解度で公平性を保つために一度視聴すると途中で動画を中断しない様にして貰った．又，認知負荷も測定しているため 1 つ目の動画の視聴で疲れが溜まっている可能性もあるため，間を空けて 2 つ目の動画の視聴を行って貰った．どちらかの動画視聴直後 GoogleForms にて認知負荷を測定するためのア

ンケートと動画の内容理解度を測定するための問題に回答してもらう．

12 人の被験者には提供する動画が被らない様にしている．また今回の実験では全てのペアを検証するための被験者を集める事が出来なかったため，提供する際に自分の考察でなるべく手法で大きく違いが出そうなペアで配っている．

第5章 実験結果

本章では 4 章の実験データを洗い出し，リアルタイム翻訳の字幕生成のタイミングについて分析を行う．その後，実験から得られた問題点について考察を行う．

5.1 内容の理解度

内容の理解度についての分析を行う．12 人の被験者の各問題の平均点数を表 4 と表 5 に示す．表 4 表 5 で記載されているのは手法別となっているピリオドと YouTube には発話翻訳での区切りがないためピリオド，YouTube という形になっている．

表 4　動画(スライドなし)の平均点数

	ピリオド	単語数 逐次翻訳	単語数 連結翻訳	カンマ 逐次翻訳	カンマ 連結翻訳	YouTube
平均点数	62.50	37.50	50.00	50.00	31.25	68.75

表 5　動画(スライドあり)の平均点数

	ピリオド	単語数 逐次翻訳	単語数 連結翻訳	カンマ 逐次翻訳	カンマ 連結翻訳	YouTube
平均点数	75.00	68.75	81.25	81.25	68.75	62.50

まず手法別の内容理解度の比較を行いたいと思う，(スライドなし)と(スライドあり)の動画で平均点数が違い，動画(スライドあり)の方が高い結果となっている．この要因として考えられる物が動画(スライドあり)は動画自体がスライドを見ながら字幕を見る物となっている．そのため，字幕を見ている時に動画自体の内容も分かるため平均的に高くなっていると思われる．(スライドなし)と(スライドあり)動画で t 検定を行った所 p 値は 0.0104 となり有意差があると分かった．そのため，動画(スライドあり)の方が動画(スライドなし)より内容を理解する点で優れているということが分かった．

手法別で見ると表 4 から動画(スライドなし)では E の手法，表 5 から動画(スライドあり)では F の手法が一番点数が近い物となっている．ただ，動画(ス

ライドあり)F の手法に関しては一人の被験者が問題文を読んでおらず回答数が一つだけとなっているため平均点数が下がっている可能性がある．そのため，動画()スライドありは B と E の手法で見るが最適だと考えている．これらから，(スライドなし)(スライドあり)の動画共に E の手法が一番理解度が低い結果となっており，E の手法はカンマ区切りの連結翻訳である．このことから，E の手法の場合，従来のリアルタイム翻訳と同様に読んでいる最中に翻訳文が一気に変わる事がある．また話者が接続詞を使うとその都度翻訳が変わったり，Finaly などの直前がピリオドで終わっていて文頭に接続詞が来る場合直前に翻訳した文が Finaly の部分でしか見る事が出来ないため重要な部分が全く読めていない可能性が高く，E の手法が低いのだと考えられる．

次に理解度の高い物を見ていく．表 4，表 5 から動画(スライドなし)は F の手法であり，動画(スライドあり)は C と D の手法となっている．動画(スライドなし)は話者が話している所に字幕があるだけの動画のため，音声とずれて字幕を表示する手法よりも既に翻訳したものを見やすくしている YouTube の動画 F の手法が一番良い結果となったと考えられる．動画(スライドあり)の場合単語数の連結翻訳である C の手法とカンマで区切りの逐次翻訳である D の手法が同率で一番高いものとなっている．この結果から C の手法は連結翻訳のため，読みにくい物となっているが 10 単語数という決められた区切り方をしているので翻訳結果が変わるとしても一定の間隔で字幕が変わるため内容が大きく変わることが少なく読み手にとって理解度が高くなったのだと考えられる．次に D の手法だが，こちらはカンマで区切りだけで読んでいる最中に文が変わる事がないため，文頭に来る接続詞以外は字幕をスムーズに全文読む事が出来ると考えられるため，読み手は全体的に内容を理解ができるためこの結果になったのではないかと思われる．また，逐次翻訳と連結翻訳での t 検定と文分割で用いている 3 つの手法カンマと単語数，カンマとピリオド，単語数とピリオドでのボンフェローニ補正を用いた t 検定を行った．その結果，逐次翻訳と連結翻訳では p 値が 0.883 となり有意差なしとなっている．次に文分割で行った物が表 6 となっている．表 6 から全てにおいて有意差なしという結果が分かった．

表 6　内容理解度に関する t 検定

組み合わせ	p 値	bonffoni
単語数とカンマ	0.883624	2.650871
カンマとピリオド	0.367059	1.101176
単語数とピリオド	0.433236	1.299708

5.2 認知負荷についての比較

次に認知負荷についての分析を行う．12 人の被験者に NASA - TLX を用いたアンケートを行った結果の各手法における認知負荷平均が表 7，表 8，表 9，表 10，表 11，表 12，表 13，表 14，表 15，表 16，表 17，表 18，となっている．

これらの結果から比較を行うと，まず認知負荷が大きい手法は動画(スライドなし)と(スライドあり)共に連結翻訳を行っている C，E の手法だと分かる．何故こうなったか考えると連結翻訳は発話の分割単位を受けるとその都度翻訳を行っているので字幕で表示するものもその都度変わっていくため既存のリアルタイム翻訳と同じで精神的な要求と時間的な要求，不満の数値が高くなっていることが挙げられる．

YouTube の動画である F の手法を除いて本研究の手法で認知負荷が少ないものは動画(スライドなし)(スライドあり)共に A，D の手法となった．この結果は A，D の手法は両方とも翻訳字幕の生成方法が逐次翻訳であり，区切り方がピ

表 7　動画(スライドなし)
ピリオド分割

ピリオド			
	評価	計	重み
精神的な要求	42.5	4	0.2665
身体的な要求	5	0	0
時間的な要求	60	4	0.266
効率	60	2	0.133
努力	50	3.5	0.233
不満	40	1.5	0.0995
総合	52.45		

表 8　動画(スライドなし)
単語数固定分割：逐次翻訳

単語数：逐次翻訳			
	評価	計	重み
精神的な要求	70	3.5	0.233
身体的な要求	5	0	0
時間的な要求	50	4	0.266
効率	90	2	0.133
努力	30	2.5	0.1665
不満	50	3	0.1995
総合	55.3		

表 9　動画(スライドなし)
単語数固定分割：連結翻訳

単語数：連結翻訳			
	評価	計	重み
精神的な要求	90	4	0.266
身体的な要求	5	1	0.0665
時間的な要求	90	4	0.2665
効率	90	3.5	0.233
努力	50	0.5	0.033
不満	70	2	0.133
総合	83.6		

表 10　動画(スライドなし)
カンマ分割：連結翻訳

カンマ：逐次翻訳			
	評価	計	重み
精神的な要求	50	3.5	0.233
身体的な要求	12.5	2	0.133
時間的な要求	70	1.5	0.1
効率	60	3.5	0.233
努力	60	3	0.1995
不満	70	1.5	0.0995
総合	52.65		

表 11　動画(スライドなし)
カンマ分割：連結翻訳

カンマ：連結翻訳			
	評価	計	重み
精神的な要求	100	4	0.2665
身体的な要求	12.5	0.5	0.033
時間的な要求	100	3	0.1995
効率	80	3.5	0.233
努力	80	1	0.0665
不満	100	3	0.1995
総合	92.6		

表 12　動画(スライドなし)
YouTube

YouTube			
	評価	計	重み
精神的な要求	30	3.5	0.233
身体的な要求	50	1	0.063
時間的な要求	40	1.5	0.1
効率	40	2.5	0.1665
努力	90	3.5	0.133
不満	40	3	0.1995
総合	49.95		

表 13　動画(スライドあり)
ピリオド分割

ピリオド			
	評価	計	重み
精神的な要求	12.5	3.5	0.233
身体的な要求	5	0	0
時間的な要求	30	3.5	0.233
効率	60	2.5	0.1665
努力	70	3	0.1995
不満	22.5	2.5	0.466
総合	36.5		

表 14　動画(スライドあり)
単語数固定分割：逐次翻訳

単語数：逐次翻訳			
	評価	計	重み
精神的な要求	60	4	0.265
身体的な要求	20	0	0
時間的な要求	60	3.5	0.233
効率	70	2	0.133
努力	70	2	0.133
不満	70	3.5	0.233
総合	65.3		

表 15　動画(スライドあり)
単語数固定分割：連結翻訳

単語数：連結翻訳			
	評価	計	重み
精神的な要求	60	4	0.2665
身体的な要求	12.5	0	0
時間的な要求	80	4	0.266
効率	60	3.5	0.233
努力	80	2	0.133
不満	60	1.5	0.0995
総合	65.95		

表 16　動画(スライドあり)
カンマ分割：逐次翻訳

カンマ：逐次翻訳			
	評価	計	重み
精神的な要求	50	2	0.133
身体的な要求	12.5	2	0.133
時間的な要求	60	3.5	0.233
効率	30	4	0.2665
努力	70	2	0.133
不満	60	1.5	0.0995
総合	49.3		

表 17　動画(スライドあり)
カンマ分割：連結翻訳

カンマ：連結翻訳			
	評価	計	重み
精神的な要求	90	4	0.266
身体的な要求	12.5	0.5	0.033
時間的な要求	90	2.5	0.1665
効率	80	1.5	0.0995
努力	42.5	2.5	0.166
不満	90	4	0.2665
総合	83		

表 18　動画(スライドあり)
YouTube

YouTube			
	評価	計	重み
精神的な要求	5	3	0.1995
身体的な要求	5	1	0.0665
時間的な要求	30	2	0.133
効率	20	4.5	0.2995
努力	40	1.5	0.1
不満	40	3	0.2
総合	26		

表 19　動画(スライドあり)と(スライドなし)の認知負荷の平均

	負荷平均
動画（スライドなし）	64.425
動画（スライドあり）	54.34167

リオド，とカンマである．

これは早い間隔で字幕が変わり続ける単語数で区切る手法は既存手法と似たような字幕の表示になるため，発話の分割単位が大きいピリオドのカンマだと基本的に字幕に表示される文が長くなっている．しかし，発話の分割単位が大きいので字幕が表示されている時間が単語数で区切る手法よりも長くなるので文が長くなったとしても字幕が表示される時間が長いため，字幕が頻繁に変わる事が少なく見やすい字幕となっているので A, D の手法は認知負荷が少ない結果となっている．

表 19 に(スライドなし)と(スライドあり)動画の認知負荷の平均を示す．動画(スライドあり)が動画(スライドなし)よりも認知負荷の平均が 18%低い事が分かる．これは字幕を見ている時にスライドも自然と目に入る事で認知負荷が高くなるのではなく被験者にとって努力と効率の尺度が下がっている．このことか

ら内容が理解しやすくなっている事が分かり，結果的に認知負荷が少なくなり内容理解度の平均点数も高くなったことが分かる．
また(スライドなし)と(スライドあり)動画，逐次翻訳と連結翻訳での t 検定と文分割で用いた 3 つの手法での単語数とカンマ，単語数とピリオド，カンマとピリオドでのボンフェローニ補正を用いた t 検定を行った．最初に(スライドなし)と(スライドあり)の動画での t 検定の結果だが，こちらの p 値は 0.0974 となっているため有意差なしと分かった．次に逐次翻訳と連結翻訳での t 検定だが，こちらの p 値は 0.0079 となっているため有意差ありと分かった．この結果から逐次翻訳の方が認知負荷より少ないことが立証された．最後に文分割での t 検定の結果が表 20 となっている．表 20 から全てのペアにおいて有意差なしということが分かった．

表 20　認知負荷に関する t 検定

組み合わせ	p 値	bonferroni
単後数とカンマ	0.865844	2.597533
単語数とピリオド	0.100551	0.301653
カンマとピリオド	0.074586	0.223757

5.3 動画別尺度の比較

5.3.1 動画(スライドなし)

動画(スライドなし)の場合は比較対象である手法 F と比べて見ると全ての手法において努力の尺度が低い数値を出している．これは動画の平均点数を見るからに F の手法は話しているタイミングで字幕が表示され見やすくなっているため問題を解く事が出来たが，他の手法では見にくい又は内容が難しいなどの点から問題を解く事を諦めてしまいこの様な結果になっていると考えられる．

5.3.2 動画(スライドあり)

動画(スライドあり)の場合手法 A だけが手法 F と比べて不満の尺度が 44%軽減されている事が分かる．これは YouTube の動画の字幕は既に翻訳したものを見やすくする様に一回に見せる字幕の表示の文を短くしているので，表示が度々変わるのでそれが被験者によっては不満と感じことがあるのではないかと考えられる．この結果から，YouTube の動画は見やすい様に工夫がされているが，表示される文が短いため，もう少し表示させる文を長くしてもよいのかも知れ

ない事が分かった.

第6章 考察

本研究ではリアルタイム翻訳の字幕生成のタイミングの分析を知るために実験を行ってきた.

6.1 文分割単位

本研究で用いた発話の分割単位はピリオド，カンマ，単語数であるがその中でピリオド，カンマが平均点数と認知負荷において高い点数，低い負荷が確認できた．このことから字幕を見る動画に関しては発話の分割単位が大きい方が次の発話の分割単位を受け取るまで時間があるので字幕を見るのに余裕ができるため，見やすくなっていることが分かった.

また，動画自体も話者だけが話している動画よりも，スライドが用いられている動画の方が平均点数が高い結果が確認出来た．このことから字幕に集中して見ていても自然とスライドが目に入るがその事で背景が度々変わるのを精神的な要求と感じず，認知負荷が逆に下がっているため，スライドがある動画で字幕を使うと平均的に認知負荷が下がる事も分かった.

6.2 翻訳字幕の生成方法

本研究で用いた翻訳字幕の生成方法における発話単位は逐次翻訳と連結翻訳であるが内容理解度の点では両方とも同じような点数を出しているためどちらが優れているか断定できないが認知負荷の点では逐次翻訳の方が認知負荷が少なく優れていると言える．連結翻訳はピリオドを認識するまで翻訳を続けるため，発話の分割単位を読み取る度に翻訳文が変わるため精神的な要求，時間的な要求，不満，の主観的な側面が高い数値が出てしまう事が分かった．以上より，読み手側にとって認知負荷を軽減する翻訳字幕の生成方法は逐次翻訳だと分かった.

第7章 おわりに

遠隔会議システムなどでの英語運用能力が低い人達を助けるためにリアルタイム翻訳の字幕生成のタイミングの分析を行う事で，今後リアルタイム音声翻訳システムを作成するために支援を行ってきた．本研究の貢献は以下の通りである．

発話の分割単位

リアルタイム翻訳の字幕生成のタイミングの最適な発話の分割単位の測定を行う．動画を視聴してもらい，内容理解度チェックと認知負荷の測定を行いその効果の検証をした．ピリオドにおいては内容理解度，認知負荷に共に数値が良い結果が出ていた．次にカンマでもピリオドについて良い結果が出ている事が分かった．この結果から発話の分割単位はピリオド，カンマで区切るのが読み手側にとって優れている事が分かった．

翻訳字幕の生成方法

翻訳字幕の生成方法の最適なのは逐次翻訳が有用であることが分かった．認知負荷において逐次翻訳が連結翻訳より認知負荷の数値が 32%軽減されている事が分かった．

動画(スライドなし)では全ての手法が比較対象である F の手法において努力の尺度が低い値が出ており，また，動画(スライドあり)では A の手法のみ F の手法より不満の尺度が低い結果が出ていることが分かった．

本研究では発話の分割単位と翻訳字幕の生成方法の点からリアルタイム翻訳の字幕生成のタイミングの分析を行った．しかし，その際翻訳を行う際のでる，翻訳を字幕に生成するずれを考慮出来ていなかった．そのため，本来よりも読みやすい字幕又は読みにくい字幕になっていた恐れがある．

今後は翻訳を出力するまでのずれも考慮して実験を行うべきである．また，F の手法を比較として置いていたが，F の手法は被験者全員に視聴してもらった方がいいと考えられる．最初に一番認知負荷が少ない YouTube 動画を視聴してもらい，その後手法を視聴する方が被験者にとって認知負荷のアンケートの際に，評定しやすいものになっているのではないかと考えられ，その結果認知負荷の測定がより公平的なものになり，数値が高い手法や低い手法が明らかになったのではないかと思われる．

このことから被験者は手法全てのペアで検証出来る様に 12 人より多い人数を

集めで行う必要がある．また，実験の内容自体もよりリアルタイム翻訳を追及するために翻訳時のずれを考慮した字幕のデザイン，認知負荷の公平性から被験者全員に 2 つの動画の提供だけではなく 3 つ目の YouTube の動画の視聴を行う事で更なる期待できるかもしれない．

謝辞

本研究を行うにあたり，熱心なご指導，ご助言を賜りました村上陽平教授に深謝申し上げます．また，普段からお世話になっている社会知能研究室の皆様と実験に協力して頂いた方々に感謝の意を表します．

参考文献

[1] 清水宏晃, Neubig, G, Sakti, S, 戸田智基, 中村哲： 同時通訳データを利用した同時音声翻訳のための訳出タイミング決定手法, 言語処理学会第 20 回年次大会 (NLP2014), pp. 294-297(2014).

[2] V.K.R. Sridhar, J. Chen, S.V. Bangalore, A. Ljolje, and R.C. Varayan : Segmentation Strategies for Streaming Speech Translation, In Proc. the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL2013), pp.230-238, 2013.

[3] 清水 宏明, Graham Neubig, Sakriani Sakti, 戸田 智樹, 中村 哲： 同時通訳データを利用した同時通訳用機械翻訳システムの構築,情報処理学会研究報告,1-5, 2013

[4] C. Fügen, A. Waibel, M. Kolss : Simultaneous Translation of Lectures and Speeches, Machine Translation 21, 209-252 (2007).

[5] 稲垣洸雄, 松村敦, 宇陀則彦： 動画における字幕デザインのパーソナライズ要素の検討, 情報知識学会誌, Vol.26, No.2, pp.239-244 (2016)

[6] 芳賀繁,水上直樹:日本語版 NASA - TLX によるメンタルワークロード測定 各種室内実験課題の困難度に対するワークロード得点の感度,人間工学, Vol.32, No.2, pp.71-79(1996).