THE JOURNAL OF ASIA TEFL
Vol. 17, No. 1, Spring 2020, 143-159
http://dx.doi.org/10.18823/asiatefl.2020.17.1.9.143

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

An Analysis of the Errors in the Auto-Generated Captions of University Commencement Speeches on YouTube

Jeong-Hwa Lee
Hansung University, Korea

Kyung-Whan Cha
Chung-Ang University, Korea

Auto-generated captions on YouTube have proven useful in helping viewers better understand the words being spoken. However, at times they fail to contain accurate captions. In these cases, they lead to confusion. The aim of this paper is to identify and analyze errors in the auto-generated captions of 20 commencement speeches on YouTube. These speeches were presented over a period of 12 years by speakers from different walks of life. The researchers selected ten male and ten female icons. Only the first 10 minutes of the speeches were utilized for this investigation. All the captioned errors were collected and analyzed. Upon completion of the analysis, it was discovered that the frequency of errors in each speech ranged between 10 and 46 cases, with an average of one error occurring about every 26 seconds. Among the different error categories, nouns record the highest number with 144 cases (31.3%). The second is verbs with 93 cases (20.2%), then prepositions with 37 cases (8.1%). Among the four subcategories, namely omission, addition, substitution, and word order, substitution recorded the highest amount of errors with 357 cases (77.6%). Furthermore, the errors were classified into two major groups. The first, involving function words, appeared in 169 cases (36.7%). The second, involving content words, appeared in 291 cases (63.3%). The results of this research suggest that a continuous development of the voice recognition software that automatically generates captions is necessary for more efficient and accurate data that will help viewers and listeners better comprehend the video contents.

Keywords: **auto-generated caption errors, YouTube, university commencement speeches, function words, content words, omission, addition, substitution, word order**

Introduction

Nowadays, learners living in EFL countries who are eager to develop their English proficiency actively use audio-visual materials on YouTube, the largest free video sharing site in the world. It is available in 91 countries in 80 different languages at present (Suh & Cho, 2019). Videos are useful learning materials in that they reproduce various characteristics of target language usage in a more communicative manner (Shin, 1998). They are able to recreate authentic language situations. Thus, they are able to provide authentic language situation. Nevertheless, listening comprehension is considered a complex continuous process (Chen, 2013). The complex vocabulary, as well as difficulty of understanding the pronunciation of native English speakers makes the learning experience challenging (Cha, 2000). Another factor to consider is the accent of the speaker. If the viewer has difficulty understanding the speech due to the accent in which it is spoken, this will affect his comprehension ability as well (Chen, 2011). This is the

one reason why captions are employed. The purpose is to help learners better understand the content. The use of these subtitles has resulted in learners experiencing a lot of improvement in their target language. The use of these subtitles has resulted in learners experiencing a lot of improvement in their target language (Ahn, 2013; Baker, 2001; Brasel & Gips, 2014; Caimi, 2006; Danan, 2004; Hinkin, Harris, & Miranda, 2014; Lee, 2017; Mitterer & McQueen, 2009; Yoon, 2018; Yu, 2013).

Markham (1989) conducted a study focusing on the effectiveness of using captioned video in the development of listening comprehension skills of ESL students belonging to various proficiency levels. The subjects were University level ESL students. The majority were from North Asian countries. The results were intriguing. Generally, it was found that those students who watched videos with captions experienced better comprehension results. But as the viewing material grew more complex, the difference was even more distinct. Also, the advanced-level students benefited just as much from the experiment as the intermediate and lower-level students. These results supported much of the previous studies, as well as studies conducted since then, in claiming that captioned videos do more to help increase comprehension levels than non-captioned videos. One such study by Koskinen, Wilson, and Jensema (1986) postulated that the simultaneous processing of both audio and text enhanced learning. A study by Guan and Ma (2018) investigated the effect of using captioning in video to improve vocabulary learning. They had two groups of high school students participating in the experiment. One group merely watched the clips, while the second group were provided captions to read while watching. They discovered that those who watched video clips containing captions experienced a greater boost in their vocabulary than those who merely watched the clips alone. This result raised discussions regarding how to make use of video effectively in classroom learning.

Another study by Teng (2019) revealed similar results as Guan and Ma. The difference was in the study subjects. Instead of high school students, he used 257 primary school students who were taking ESL classes. He divided them into six groups. Each group viewed the same video under different conditions. In the post-test administered afterward, the group that viewed the fully-captioned version of the video performed significantly better than the others. He discovered that fully captioned videos are better at improving the comprehension of ESL primary school learners than keyword captioned or uncaptioned videos, with the student's English proficiency level being an influential factor.

Ideally, subtitles are encoded by the producer of the video before it is published. It is done manually, especially in cases where the video is lengthy and has a wide variety of characters. This often demands considerable time and effort (Johnson, 2014). This is the reason why the majority of videos on YouTube do not contain preset subtitles. This number even at times includes content published by officially-recognized sources. But with the advent of a number of new functions, such as voice command and assistance, more precise recognition technology has been developed (Han, Yang, & Kim, 2019). One of the benefited areas is online video-sharing platforms such as YouTube. This has made the use of subtitles quite ubiquitous. This is seen in the availability of auto-generated captions in videos that do not contain preset captions. This feature is present in most devices that can stream videos, including tablets, PCs, and smartphones.

The speech recognition technology utilized is designed to recognize human speech and convert it into text. It compares the input speech data with the entire speech and language database in order to generate the appropriate text in real-time compared with typical physical interfaces such as the mouse and keyboard, this innovative technology works by automatically reading the human voice. Moreover, input speeds through speech recognition programs are typically two to three times faster than conventional physical interfaces (Carey, 2016).

Even though speech recognition technology has experienced advancement since its arrival, users still experience some degree of difficulty due to recognition errors (Partron, 2016). It has a high accuracy rate when it came to single syllables and short words and phrases, however, when used in a noisy environment where different voices overlap; the accuracy rate greatly diminishes (Lee, Yang, & Kwon, 2001). YouTube also provides the capability to translate the auto-generated subtitles into other languages, but the error rate rises significantly.

To conduct this study, the researchers chose to use university commencement speeches. Because of the formal setting in which these speeches were given, the pronunciation and grammar were on par, and thus qualified as good research material to use. These well-known individuals including writers, business leaders, presidents, and star entertainers, are also influential in the direction of life young graduates take, which is revealed in their manner of conversation (Choi, 2016).

We expect that the results of this study will lead to the development of a more accurate captioning program on YouTube, which will generate an exact copy of the speaker's words. The objective of the study seeks to answer the following research questions:

1. What are the error rates according to the 10 error classifications and their sub-categories?
2. What are the frequencies of errors that fall under function words and content words?
3. How do the rates of the identified captions errors in the 20 speeches according to genders?

Literature Review

Auto-generated Caption Errors

In this section the researchers looked at previous studies dealing with auto-generated caption errors and machine translation. Currently, studies relating to auto-generated caption errors are few in number, compared to studies relating to machine translation. Captions, now a common feature in media, used to be nonexistent not too long ago. They are generally defined as a textual representation of spoken words presented in video (Gernsbacher, 2017; Learning Center, 2019). There are two types of captioning, namely open captions and closed captions. While closed captions are adjustable within the player and can be switched on or off by the user, open captions are fixed into the video and cannot be tampered with (Clossen, 2014; Do-IT, 2019). Auto-generated captions are always closed captions. Subtitles, on the other hand, are similar to captions except that they are translations in other languages. They are intended for people who do not speak or understand the language in the video (Flynn, 2016). YouTube provides both captioning and subtitle services. It is worth noting however, that subtitles that are auto-generated often contain more errors than captions that are auto-generated. Most of the available studies dealing with these two features emphasize their use in education. The focus of this research will be the errors found in the auto-generated captions of web-based videos. These errors will be classified, and studied closely to determine how they influence the effective use of video in learning English.

In Partron's research (2016) dealing with auto-generated YouTube captions that meet the needs of deaf students, he believed that providing captions for videos viewed in online courses is a topic of interest for institutions of higher education. His research investigated a specific type of video, weekly news updates uploaded by professors teaching online classes, in order to determine whether automatic captions are accurate enough to fulfill the learning needs of deaf students. Out of the total of 68 minutes of captioned videos that were inspected, 525 phrase-level errors were detected. These translated to an average of 7.7 phrase errors every minute. The findings reveal that the auto-generated captioning systems existing today do not meet the accuracy levels necessary for hearing-impaired learners.

A study by Tatman (2016) sought to investigate the accuracy of the automatically-generated captioning system used by YouTube when both genders and five English dialects are involved. Eight men and eight women were chosen. The dialects selected were California, New Zealand, New England, Georgia, and Scotland. In carrying out the experiment, the accent tag challenge was employed, which provided the opportunity for the identification of the speakers' various language backgrounds. The results revealed clear differences across both gender and dialect. There was a higher error rate for women than men. Regarding dialect, Scottish speakers had a worse performance than the other five, with New Zealand and Georgia close behind. This study strengthens the point that the socio-linguistic level of the speaker may affect their ability to use speech recognition accurately.

Doherty and Kruger (2018) looked into the quality of human and machine-generated captioning systems. They examined the various industry standards developed to guarantee that viewers are presented with high-quality subtitling and captioning regardless of language or disability. Although not all viewers are aware of subtitles and captions, and even ignore it altogether, yet studies make it clear that everyone can benefit from well-designed subtitles and captions. In an age where the AVT (audio-visual translation) technology continues to gain popularity, there has never been a better opportunity for meaningful discussion and collaboration between language systems and AVT across all sectors. Involvement of language technologies and raising awareness regarding their strengths and limitations are bound to be of benefit to AVT users, especially as more advanced resources are available online (e.g., Igareda & Matamala, 2011).

Machine Translation Errors

Machine Translation is an automatic conversion of text or speech between languages. The auto-generated captioning system that YouTube provides makes use of this technology. Just like voice recognition systems, it is not accurate or free from errors (Afshin & Alaeddini, 2016; Ghasemi & Hasheminan, 2016; Harmeier & Guillou, 2018; Park, 2017; Vidhayasai et al., 2015). However, it is widely used as a useful educational tool in the learning of foreign languages (Bahri & Mahadi, 2016; Lee & Cha, 2019a; Murtisari et al., 2019). Vilar et al. (2006) grouped machine translation errors into five major categories, word order, punctuation, incorrect words, missing vocabulary, and indefinite vocabulary. They concentrated on three translation objectives: English to Spanish, Spanish to English, and Chinese to English. While examining the errors, they concluded that the most important error type was language-dependent errors. This refers to the generating of verb tenses necessary for translation of English into Spanish, or the word order used when translating from Chinese into English. In a research conducted by Briggs (2018), he discovered that the error rate of web-based machine translation systems was significant when used as an assistive learning tool for Korean university students. In a survey involving 80 students who regularly used it, most of them reported a low degree of trust in the data output. Furthermore, there was a disparity between the perceived value of, and the actual effectiveness of such tools. Li, Graesser, and Cai (2014) also point out that “machine translation fails in grammar accuracy, cases with complex syntax, semantics, and practical structures.” In a study comparing Google Translate with Human Translate, Aslerasoul and Abbasian (2015) discover that the efficacy of Google Translate was somewhat comparable to human translate when relating to specific fields such as physics and politics. Lee and Cha (2019b) analyze machine translation errors based on eleven categories. Using Google Translate, they translated 30 newspaper articles published in the Health Chosun in 2017 dealing with food and health from Korean into English. The results revealed that of the 543 sentences taken from the 30 articles, 365 (67.2%) contained translation errors. Only 178 (32.8%) were accurately expressed.

Although auto-generated errors are often found, they vary depending on whether captions or subtitles are being considered. Machine translation errors have to do with voice recognition systems, where the speaker’s words are converted into text presented in the same language. This captioning feature often results in a lot of errors because the program is simply trying to match each spoken word with the corresponding term in the database that most resembles it.

Method

Data Collection

For this study, 20 commencement speeches were selected from YouTube by the researchers. Each speech was presented by a different individual. These individuals are well-known global icons. Their occupations vary, as listed below in Table 1. Four hold high offices in the country, eight are big business

executives, and another eight are celebrities and cultural influencers. The majority are American because the colleges and universities where the speeches were given are all located in the United States. In the selection process, no preference was given to race or nationality. It was intended, however, that each gender be equally represented: 10 males and 10 females. All the speeches were given within a 12-year span, from 2007 to 2019.

TABLE 1
List of Selected Commencement Speakers

N	Gender	Speaker	Title or position	Location of speech	Year
1	M	Barack Obama	Former president of the United States of America	Howard University	2016
2	M	Bill Gates	Founder of Microsoft	Harvard University	2007
3	M	Conan O'Brien	Talk-show host and comedian	Dartmouth College	2011
4	M	Eric Schmidt	Former CEO of Google	Boston University	2012
5	M	Jeff Bezos	Founder and CEO of Amazon	Princeton University	2019
6	M	Mark Zuckerberg	Co-founder and CEO of Facebook	Harvard University	2017
7	M	Robert De Niro	Actor, producer, & director	New York University	2015
8	M	Steve Jobs	Co-founder and former CEO of Apple	Stanford University	2005
9	M	Tim Cook	CEO of Apple	Stanford University	2019
10	M	William McRaven	Retired US Navy admiral	Texas University	2014
11	F	Hillary R. Clinton	Former US secretary of state	Yale University	2018
12	F	Joan K. Rowling	Author of the Harry Potter book series	Harvard University	2008
13	F	Mary Barra	Chair and CEO of General Motors	Michigan University	2014
14	F	Meryl Streep	Academy award winning actress	Barnard University	2010
15	F	Michelle Obama	Former first lady of the United States of America	City University of New York	2016
16	F	Mindy Kaling	Comedian, actress, and writer	Dartmouth College	2018
17	F	Natalie Portman	Actress and filmmaker	Harvard University	2015
18	F	Oprah Winfrey	Talk show host and philanthropist	Harvard University	2013
19	F	Savannah Guthrie	Journalist and attorney	Georgetown University	2019
20	F	Sheryl Sandberg	COO of Facebook	Massachusetts Institute of Technology	2018

Data Analysis

The data were analyzed with IBM SPSS (23.0) software. Once the assessment stage was done, the errors were classified into 10 categories, namely nouns, pronouns, verbs, auxiliary verbs, articles, prepositions, adjectives, adverbs, conjunctions and uncategorized errors which include contractions and exclamation marks. As shown in Table 2, the errors of the 10 categories are further classified into four subcategories. Those four subcategories are errors of omission, addition, substitution, and word order, resulting in a classification system involving 40 error types. Errors of omission occur when the auto-generated captions fail to record a word or phrase that the speaker said, while errors of addition occur when certain words or phrases appear in the captions that were never spoken in the video. Substitution errors include misspelling, proper nouns, and countable and uncountable nouns, etc. Errors involving word orders occur when the arrangement of words that appear in the captions are not synchronized with the arrangement of words spoken. All four classifications cover every type of error that was assessed by the researchers.

TABLE 2
The Error Classification for Auto-generated Caption Errors

N	The Error types	Sub - Categories
1	Noun	Omission, Addition, Substitution, Word Order
2	Pronoun	Omission, Addition, Substitution, Word Order
3	Verb	Omission, Addition, Substitution, Word Order
4	Auxiliary Verb	Omission, Addition, Substitution, Word Order
5	Article	Omission, Addition, Substitution, Word Order
6	Preposition	Omission, Addition, Substitution, Word Order
7	Adjective	Omission, Addition, Substitution, Word Order
8	Adverb	Omission, Addition, Substitution, Word Order
9	Conjunction	Omission, Addition, Substitution, Word Order
10	Uncategorized: Contraction & Exclamation	Omission, Addition, Substitution, Word Order

As shown in Table 3, categories that fall under function words include pronouns, articles, prepositions, conjunctions, exclamations, contractions involving auxiliary verbs, and contractions involving pronouns. Categories that fall under content words include nouns, verbs, adjectives, adverbs, and contractions involving negatives.

TABLE 3
Classification of Caption Errors Based on Function and Content Words

The Error types	Error Categories
Function words	pronouns, articles, prepositions, conjunctions, exclamations, contractions involving auxiliary verbs, contractions involving pronouns
Content words	nouns, verbs, adjectives, adverbs, contractions involving negatives

Results

Auto–Generated Caption Errors Based on 10 Categories and Four Sub-Categories

Frequency of auto–generated caption errors based on 10 categories

The frequency rates of errors among the 10 categories are presented in Table 4. Out of all 10, nouns record the highest number of errors with 144 cases (31.3%). The second is verbs with 93 cases (20.2%). Third comes prepositions with 37 cases (8.1), then uncategorized errors with 36 (7.8%), which includes exclamation and contraction errors. Next comes pronouns with 34 cases (7.4%), adjectives with 33 (7.2%), articles with 30 (6.5%), conjunctions with 24 (5.2%), adverbs with 17 (3.7%), and lastly, auxiliary verbs with 12 cases (2.6%), the lowest number of errors recorded.

TABLE 4
The Results of Auto-generated Caption Errors Based on the 10 Categories

Types of Errors	Frequency	Percent	Cumulative Percent
Noun	144	31.3	31.3
Pronoun	34	7.4	38.7
Verb	93	20.2	58.9
Auxiliary Verb	12	2.6	61.5
Article	30	6.5	68.0
Preposition	37	8.1	76.1
Adjective	33	7.2	83.3
Adverb	17	3.7	87.0
Conjunction	24	5.2	92.2
Uncategorized contraction & exclamation	36	7.8	100.0
Total	460	100.0	

Frequency of addition, omission, substitution and ordering error

Table 5 shows the frequency of errors relating to the four subcategories, namely addition, omission, substitution and ordering. Among the four, substitution recorded the highest amount of errors with 357 cases (77.6%). Omission came second with 84 cases (18.3%), and addition with 19 cases (4.1%). There are no errors involving word order. These results suggest that the speech recognition program is merely following the pattern of speech. It transcribes the speaker's words directly, without considering punctuation, and without analyzing the relation between different phrases and sentences.

TABLE 5
Error Frequency Rates Based on the Four Subcategories

Types of Errors	Frequency	Percent	Cumulative Percent
Omission	84	18.3	18.3
Addition	19	4.1	22.4
Substitution	357	77.6	100.0
Word order	0	0	100.0
Total	460	100.0	

Function Word and Content Word Errors

Frequency of function word and content word errors

In order to identify the different types of errors in general, two groups were formed, content words and function words (Table 6). Function words recorded 169 errors (36.7%). They are pronouns, articles, prepositions, conjunctions, exclamations, contractions dealing with auxiliary verbs, and contractions dealing with pronouns. Content words, on the other hand, recorded 291 errors (63.3%). They are nouns, verbs, adjectives, adverbs, and negative contractions.

TABLE 6
Frequency of Function Word and Content Word Errors

Types of Errors	Frequency	Percent	Cumulative Percent
Function Words Errors	169	36.7	36.7
Content Words Errors	291	63.3	100.0
Total	460	100	

Examples of content word errors

This section presents several examples of errors involving content words. Quotes from seven speakers are presented in Table 7. In the first example involving nouns, Natalie Portman repeats the phrase "city steps!" four times. The voice recognition program detects each repetition differently, as shown in the table below. An example of a verb error is Sheryl Sandberg using the past tense of the word "turn" in her speech, but it is displayed in the present form instead. When it comes to adjective errors, Mindy Kaling mentions the word "many," but it is captioned as "mini, " which is completely different. With adverb errors, the word "enormously" is divided into two separate words, "enormous" and "lee." This disrupts the coherence of the sentence. The fourth category is contractions involving negatives. Three different examples are given here. The first error was detected in Savannah Guthrie's speech. Her phrase "stick to what's safe" is captioned as "stick to wet safe." The second involves a possessive noun. Instead of parent's, the caption reads "parents." The apostrophe is excluded, thus changing the meaning entirely. The third example is a typical negative contraction. Instead of isn't, the caption reads "is an." This also changes the intended meaning.

TABLE 7
Examples of Content Word Errors

Type of Content Word	Original spoken text	Auto-generated captions	Commencement speaker	Time of occurrence
Nouns	City steps! City steps city steps city steps	city stuff city stuff city steps cities death	Natalie Portman	8:55
Verbs	when I need advice I turned to Mark Zuckerberg	when I need advice I turn to Mark Zuckerberg	Sheryl Sandberg	2:11
Adjectives	we will be forced to be many experts on dr. Seuss	we will be forced to be mini experts on dr. Seuss	Mindy Kaling	0:37
Adverbs	reflecting on her speech has help me enormously in writing	reflecting on her speech has help me enormous Lee in writing	Joan K. Rowling	1:45
Contractions involving negatives	stick to what's safe I'll tell you	stick to wet safe I'll tell you	Savannah Guthrie	7:02
	my working-class parent's saving were	my working-class parents saving were	Steve Jobs 2:42	2:42
	in general advice isn't actually an effective way	in general advice is an actually an effective way	Mindy Kaling	7:32

Examples of function word errors

This section presents six examples of function word errors. The first is pronouns. In Mindy Kaling's speech, she mentions the phrase "you just had to sit there," but it is captioned as "he just had to sit there." This changes the subject of the sentence. An example of an error involving articles is found in Tim Cook's phrase "gotten lucky with the wind," which is detected as "lucky with a wind." A slight change in meaning, but still significant. A phrase spoken by Oprah Winfrey is given as an example of a preposition error. She says "by and by," which means later. But this is captioned as "bye-bye." The fourth example involves conjunctions. Barack Obama's phrase "and stay with me now" is detected as "then stay with me now." Although the meaning isn't significantly affected, this mistake committed by the speech recognition program must not be ignored. In the column titled "time of occurrence," this error is seen to take place at the 18:05 mark. This is due to the starting time of the speech beginning 8:07 minutes into the video.

The last category is contractions involving both pronouns and exclamations. The first involving pronouns, is found in the phrase "they're all be working," which is detected from Robert De Niro's words "they'll all be working." This means that the speech recognition program can't follow through with the grammar rule in the sentence. It instead deals with individual words, looking for the closest sounding

word. The second part of this error category involves an exclamatory word. In Michelle Obama's speech, she says "whoa!" This is captioned as "we'll," completely changing the expression intended to be made.

TABLE 8
Examples of Function Word Errors

Type of Function Word	Original spoken text	Captioned phrase	Commencement Speaker	Time of occurrence
Pronouns	you were out of luck you just had to sit there	you were out of luck he just had to sit there	Mindy Kaling	5:30
Articles	we must have gotten lucky with the wind kidding aside	we must have gotten lucky with a wind kidding aside	Tim Cook	2:35
Prepositions	it's by and by when the morning comes	it's bye-bye when the morning comes	Oprah Winfrey	6:53
Conjunctions	the world is better and stay with me now race	18:05 the world is better then stay with me now race	Barack Obama	18:05 (Starting time 8:07)
Contractions Involving pronouns & exclamations	they'll all be working shitty jobs lousy pay different languages	they're all be working shitty jobs lousy pay different languages we'll	Robert De Niro	2:02
	whoa just stop there you	just stop there you	Michelle Obama	3:48

Frequency Rates of Auto-generated Caption Errors as Recorded from the 20 Commencement Speeches

Number of auto-generated caption errors recorded from the 20 speeches

This section describes the number of auto-generated caption errors as identified in 20 commencement speeches. Out of the 20 commencement speakers represented in Figure 1, 10 were male, and 10 were female. A total of 460 errors were recorded from the entire 200 minutes of speeches. The frequency of these errors ranged between 10 cases and 46 cases, and there was an average of one error occurring every 26 seconds. It is interesting to notice that the frequency of errors assessed greatly varied among the selected speakers. Among the male speakers, four stand out when assessing the errors. The first is Tim Cook, whose speech recorded only 10 errors, the least out of all 20 speakers. Steve Jobs' speech came second with only 11 errors. Eric Schmidt's speech, however, recorded a whopping 40 errors, while Robert De Niro was close behind with 36 errors. Among female speakers, the least amount of errors was recorded in Savannah Guthrie's speech, while J. K. Rowling's recorded only 12 errors. On the other hand, Mindy Kaling's speech with 46 words, accounted for the highest amount of errors. The second highest was Michelle Obama's speech with 36 errors.

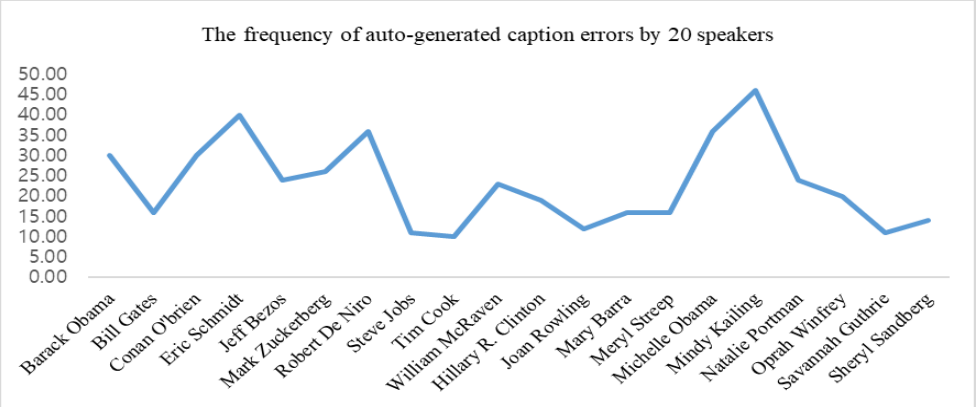

Figure 1. The number of auto-generated caption errors according to the 20 speakers.

The comparison of auto-generated caption errors according to genders

Table 9 shows the number of auto-generated caption errors according to genders. A majority of the auto-generated caption errors assessed belonged to the male speakers, accounting for 246 word errors (53.5) combined. The female speakers, however, represented only 214 word errors (46.5%). This seems to indicate that male speakers are slightly more prone to committing more errors, but it is important to bear in mind that some of the errors recorded were a result of Google's speech recognition program.

TABLE 9
The Number of Auto-generated Caption Errors According to Gender

Types of Errors	Number	Frequency	Percent	Cumulative Percent
Male	10	246	53.5	53.5
Female	10	214	46.5	100.0
Total	20	460	100.0	

Discussion and Implication

This section deals with the implications of the results of the study based on the three research questions. It is presented in three parts, each part seeking to address an individual area relating to the auto-generated caption errors on YouTube.

Relating to the 10 Categories and Four Sub-Categories

According to the study, the automatic captioning errors identified in the commencement speeches were found to involve all 10 error classifications. Noun errors accounted for the highest portion with 144 (31.3%), followed by verbs with 93 (20.2%), and the remaining eight categories recorded less than 10% each. Nouns and verbs are the main components of sentences. Due to their importance, mistakes that involve nouns and verbs result in much confusion for viewers. According to the assessment, there were many errors that had to do with spelling and the singular/plural form of nouns, and there were many errors involving verb tense. While these voice recognition systems present text in a simplified manner, words spoken in an indistinct voice are often inferred as errors being made due to the machine's poor recognition of its contents.

Furthermore, out of the four subcategories of errors, substitution was the highest with 357 out of 460 errors recorded. This often happens because the machine misinterprets the speakers' words differently. No errors were found to do with word order. These findings seem to indicate that the voice recognition program is simply rendering the words spoken. It directly transcribes the speaker's words, without regard for punctuation, or the relation between various phrases and sentences.

Relating to Function Words and Content Words

The results of the error assessment shows that content errors tend to be committed more frequently than function errors. Content words cover essential parts of speech in a sentence, so mistakes made often result in confusion for viewers.

This makes it clear that current voice recognition systems need a more advanced algorithm that makes it possible for them to comprehend the sentence as a whole. The examples presented above reveal how the program seeks to translate each word individually. When this is the case, mistakes are bound to be made because the speaker will not always pronounce the same word in the same way. An example is found in Natalie Portman's speech, where she repeats the phrase "city steps!" four times, but the voice recognition program detects each repetition differently, "city stuff, city stuff, city steps, cities death."

There are even cases where the program dissects the word and interprets each syllable differently. An example is “enormously” captioned as “enormous” and “lee.”

Function words, although they may not play a big role in the sentence as content words do, can still create confusion if interpreted wrongly. In one example in Barack Obama’s speech, he mentions the phrase “and stay with me now,” which is interpreted as “then stay with me now.” This is a minor error but it affects to get the idea of the sentence. But, it still must be addressed. Another example is seen in Robert De Niro’s speech where he says “they’ll all be working,” which is interpreted as “they’re all be working.”

Relating to the Frequency Rates of Each of the 20 Commencement Speeches

It is interesting to notice that the frequency of errors assessed greatly varied among the selected speakers. The reason for such a wide disparity of errors among 20 speakers may be due to several factors, one being accent. Individuals whose jobs often require them to address large groups of people consisting of different demographics tend to develop an indistinct accent, which is described as an accent that cannot be said to originate from any specific region. It was also noticed that the speaker’s enunciation ability to enunciate played a big role. Examples are speakers such as Tim Cook, Steve Jobs, Savannah Guthrie, and J. K. Rowling who are known to speak clearly when they give formal speeches. This makes it easy for the speech recognition program to detect their words. It is noteworthy to observe that the majority of speakers with low error rates are corporate executives and authors. This was not the case with celebrities like Mindy Kaling and Robert De Niro, or Oprah Winfrey and Conan O’Brien, all of whom are well-known icons in show business. The speeches of those four speakers gave all recorded significant amounts of errors. The Obamas, while neither comedians nor actors, also recorded a significant amount of errors. This may be due to their manner of speech.

Another factor that influences error rates is the background of the speaker. Mindy Kaling, who is of Indian ancestry, and has lived most of her life in Southern California. Her speech is often fast-paced and expressive in tone, as is expected of celebrities working in that part of the country. This manner of speaking often results in a high rate of caption errors.

A third factor that bears mentioning is venue. Of the twenty commencement speakers, three delivered their speeches indoors: Hillary Clinton, Jeff Bezos, and Robert De Niro. The other 17 speakers delivered their speeches outdoors. At first, this would seem to be a determining factor in the amount of errors assessed. But the results appear to be uninfluenced by the venue of the speech. The three indoor speeches mentioned recorded 24, 36, and 19 errors respectively. These results do not differ in any significant way from the other speeches.

Furthermore, the errors assessed showed a discrepancy between male and female speakers. In contrast with the results of Tatman’s (2016) study, the error rate of speeches given by males amounted to 53.47%, or a total of 246 word errors, while speeches given by females amounted to 46.52%, or 214 word errors. The difference is not that significant, but still bears mention. The reasons for this vary depending on certain factors such as profession, accent, race, and personality.

Summary and Limitations

This paper investigated the errors of auto-generated English captions in university commencement speeches provided on YouTube. This is a research topic that has been rarely addressed, but which has academic significance due to the prevalent use of media in learning. This research sought to answer three questions. The first question involved the rate of errors assessed according to the 10 classifications and subcategories, the second dealt with the error frequencies of function and content words, and the third was intended to determine how the errors of different speakers’ genders compared. The results of the research conducted were interesting. It was discovered that a total of 460 errors were recorded from the entire 200 minutes of speeches, with an average of one error occurring about every 23 seconds. The

highest category involved nouns, followed by verbs, then prepositions. These errors were further classified into four subcategories, of which substitution, at 77%, had the highest number of errors. This mainly involved the replacing of words with similar-sounding ones.

The data gathered from the investigation revealed that errors involving content words (63.3%) were more frequent in occurrence than those dealing with function words (36.7%). Different speeches recorded different amounts of errors. Speeches given by male speakers amounted to 53.5% of errors, and those given by females with 46.5%. Even when working within the same language, there are complex variables that the speech recognition software has to consider, such as the speaker's gender, age, and style of pronunciation, as well as moments where a word or phrase is pronounced alone and when it is pronounced as part of a sentence. The rapidity of speech is another variable that has been found to greatly diminish data accuracy. For these reasons, the algorithm of voice recognition systems are seen to be quite complex. This is a research topic that has been rarely addressed, but which has academic significance due of the prevalent use of media in learning. The results of this research will be of great help for future studies. It is the hope of the researchers that an improved captioning system will be developed that will provide all YouTube users with a better viewing experience. More than just raising awareness, the researchers believe that a better understanding of current captioning systems will enable language instructors to make good use of these features in their teaching models. For instance, the students can use the auto-generated subtitles as an aid when listening, speaking, reading, and writing activities about the video content. They can also compare the official script with the auto-generated text to identify errors so they can avoid them in their studying. Because these machine-generated captions do not contain punctuation, and sometimes match the wrong words, the student can go through the text, and practice making the necessary corrections.

The limitations of this study are as follows. In terms of sample size, the researchers dealt with 20 commencement speeches given in various American universities and colleges. All these speeches were taken from YouTube. The 20 speakers chosen in this study are all globally influential icons. Ten are male, and 10 are female. This study only dealt with auto-generated subtitling errors in 20 commencement speeches given in American universities. In the future, we hope more research utilizing a wider variety of YouTube videos will be conducted relating to auto-generated subtitling errors, with a focus on other variables, such as different phonetics and accents. We understand that it would serve a more comprehensive study for different nationalities and backgrounds to be considered. In a world where English is often contextualized, it is important to recognize that there are many varieties of English, none of which are superior or inferior to another. They should all be equally recognized and studied.

Acknowledgments

The authors appreciate the reviewers for their valuable comments and useful guidelines in the development of this paper.

The Authors

Jeong-Hwa Lee, Ph. D. is an instructor in the Department of Liberal Arts and Sciences at Hansung University in Seoul, Korea. She received her doctoral degree in English Education from Chung-Ang University in Seoul. Her research interests include error analysis in machine translation and auto-generated subtitles, and e-learning.

Department of Liberal Arts and Science, Hansung University
116 Samseongyoro 16gil, Seongbuk-gu, Seoul, Korea
Email: 2019jhlee@naver.com

Kyung-Whan Cha, Ph. D. (corresponding author) is a professor in the Department of English Education at Chung-Ang University, Seoul, Korea. He received his doctoral degree in TESL from the University of Kansas, and he has been teaching at Chung-Ang University since 1988. His primary research interests are L2 listening and its accompanying mechanisms.

Department of English Education
Chung-Ang University
84 Heukseok-ro, Dongjak-gu, Seoul, Korea
Email: kwcha@cau.ac.kr

References

Ahn, M. (2013). The impact of subtitled online video clips on incidental vocabulary learning. *STEM Journal*, *14*(2), 135-151.

Afshin, H., & Alaeddini, M. A. (2016). A contrastive analysis of machine translation (Google Translate) and human translation: Efficacy in translating verb tense from English to Persian. *Mediterranean Journal of Social Sciences, 7*(4), 40-48.

Aslerasouli, P., & Abbasian, G. R. (2015). Comparison of Google online translation and human translation with regard to soft vs. hard science texts. *Journal of Applied Linguistics and Language Research, 2*(3), 169-184.

Bahri, H., & Mahadi, T. S. T. (2016). Google Translate as a supplementary tool for learning Malay: A case study at Universiti Sains Malay. *Advances in Language and Literary Studies, 7*(3), 161-167.

Baker, M. (2001). *Routledge encyclopedia of translation studies*. London, UK: Routledge.

Brasel S. A., & Gips J. (2014). Enhancing television advertising: Same-language subtitles can improve brand recall, verbal memory, and behavioral intent. *Journal of the Academy of Marketing Science*, *42*(3), 322-336.

Briggs, N. (2018). Neural machine translation tools in the language learning classroom: Students' use, perceptions, and analyses. *The JALT CALL Journal, 14*(1), 23-24.

Caimi, A. (2006). Audiovisual translation and language learning: The promotion of intralingual subtitles. *The Journal of Specialized Translation, 6*, 85-97.

Carey, B. (2016, August 24). Smartphone speech recognition can write text messages there times faster than human typing. Retrieved from https://news.stanford.edu/2016/08/24/stanford- study- speech -recognition-faster-texting/

Cha, K. W. (2000). A study of impediments in listening to English news broadcasts. *English Teaching*, *55*(1), 101-225.

Chen, A. H. (2013). EFL listeners' strategy development and listening problems: A process-based study. *The Journal of Asia TEFL, 10(3*), 81-101.

Chen, H. C. (2011). Judgments of intelligibility and foreign accent by listeners of different language backgrounds. *The Journal of Asia TEFL, 8(*4), 61-83.

Choi, Y. H. (2016). *A commencement address from the American college of honorary*. Seoul, Korea: Jonghap Books.

Clossen, A. (2014). Beyond the letter of the law: Accessibility, universal design, and human centered design in video tutorials. *Pennsylvania Libraries: Research & Practice, 2*(1), 27-37

Danan, M. (2004). Captioning and subtitling: Undervalued language learning strategies. *Erudite, 49*(1), 67-77.

Doherty, S., & Kruger, J. L. (2018). Assessing quality in human- and machine-generated subtitles and captions. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation quality assessment* (pp. 179-197). Cham, Switzerland: Springer.

Do-IT. (2019). What is the difference between open and closed captioning? Retrieved from https://www.washington.edu/doit/what-difference-between-open-and-closed-captioning

Flynn, N. (2016, December 9). Captions and subtitles: Do you know the difference? Retrieved from https://cielo24.com/2016/12/captions-and-subtitles-difference/

Gemsbacher, M. A. (2017). Video captions benefit everyone. *Policy Insights Behave Science, 2*(1), 195-202.

Ghasemi, H., & Hashemian, M. (2016). A comparative study of Google translate translations: An error analysis of English-to-Persian and Persian-to-English translations. *English Language Teaching*, *9*(3), 13-17.

Guan, C., & Ma, Q. (2018). *Did we overlook the visual context in bilingually subtitled movies?* Paper presented at the 16th Asia TEFL 1st MAAL & 6th HAAL 2018 International Conference: English Language Teaching in the Changing Glocalised World: Research and Praxis. University of Macau, Macau, China.

Han, C. L., Yang, Y. S., & Kim, K. H. (2019). Research on developing a serviceable speaking program based on artificial intelligence voice recognition. *Primary English Education, 25*(2). 109-125.

Hardmeier, C., & Guillou, L. (2018, August 30). Pronoun translation in English-French machine translation: An analysis of error types. *Computer Science.* Retrieved from https://arxiv.org/ pdf/ 1808.10196v1

Hinkin, M. P., Harris, R. J., & Miranda, A. T. (2014). Verbal redundancy aids memory for filmed entertainment dialogue. *The Journal of Psychology, 148* (2), 161-176.

Igareda, P., & Matamala, A. (2011). Developing a learning platform for AVT: Challenges and solutions. *JoSTrans, 16,* 145-162.

Johnson, A. (2014). *Video captioning policy and compliance at the University of Minnesota at Duluth* (Unpublished master's thesis). University of Minnesota, Duluth, Minnesota.

Koskinen, P. S., Wilson, R. M., & Jensema, C. J. (1986). Closed-captioned television: A new technology for enhancing reading skills of learning disabled students. *Spectrum, 4*(2), 9-13.

Learning Center. (2019). Caption it yourself. Described and captioned media program. Retrieved from https://www.dcmp.org/ciy

Lee, G. S., Yang, S. L., & Kwon, Y. H. (2001). *Speech recognition*. Hanyang University Press.

Lee, J. H., & Cha, K. W. (2019a). A study on the effectiveness of machine translators for university freshmen in translating Korean writing into English. *Journal of Learner-Centered Curriculum and Instruction, 19*(8), 155-180.

Lee, J. H., & Cha, K. W. (2019b). An analysis of Korean-English translation errors in Google Translate. *The Journal of Linguistic Science, 89,* 221-257.

Lee, J. Y. (2017). Effects of using subtitles of video contents on L2 learners listening and vocabulary development: A meta-analysis. *Korean Journal of Applied Linguistics, 33*(2), 137-158.

Li, H., Graesser, A. C., & Cai, Z. (2014, May). *Comparison of Google translation with human translation.* In FLAIRS Conference.

Markham, P. L. (1989). The effects of captioned television videotapes on the listening comprehension of beginning, intermediate, and advanced ESL students. *Educational Technology, 29*(10), 38-41.

Mitterer, H., & McQueen, J. (2009, November 11). Foreign subtitles help but native-language subtitles harm foreign speech perception. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC2775720/

Murtisari, E. T., Widiningrum, R., Branata, J., & Susanto, R. D. (2019). Google translate in language learning: Indonesian EFL students' attitudes. *The Journal of Asia TEFL, 16*(3), 978-986.

Park, O. S. (2017). Error analysis according to the typological characteristics of source text in Korean-English machine translation. *The Journal of Society for Humanities Studies in East Asia, 41*, 155-183.

Parton, B. (2016). Video captions for online courses: Do you tube's auto-generated captions meet deaf students' needs?. *Journal of Open, Flexible, and Distance Learning, 20*(1), 8-18.

Ranchal, R., Doughty, T. T., Guo, Y., Bain, K., Martin, H., Robinson, J. P., & Duerstock, B. S. (2013). Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE*

Transactions on Learning Technologies, 6(4), 299-311.

Shin, D. L. (1998). Using videotaped lectures for testing academic listening proficiency. *International Journal of Listening, 12*(1), 57-80.

Suh, J. Y., & Cho, S. E. (2019). Translation of YouTube K-beauty contents. *The Journal of Translation Studies, 20*(1), 127-155.

Tatman, R. (2016). Speaker dialect is a necessary feature to model perceptual accent adaptation in humans. *4th Pacific Northwest Regional NLP Workshop: NW-NLP 2016.*

Teng, F. (2019). Maximizing the potential of captions for primary school ESL students' comprehension of English-language videos. *Computer Assisted Language Learning, 32*(7), 665-691.

Yoon, S. D. (2018). Learners' prior learning experience of subtitles and its effect on reading competence and listening competence: Using subtitles for the newsroom season 1. *STEM Journal, 19*(3), 65-82.

Yu, D. H. (2013). The effectiveness of movie subtitles in activating output-first activities: Using you again. *STEM Journal, 14(*1), 119-136.

Vidhaysasai, T., Keyuravoung, S., & Bunsom, T. (2015). Investigating the use of Google Translate in "terms and conditions" in an airline's official website: Errors and implications. *Journal of Language Teaching and Learning in Thailand, 49*, 137-169.

Vilar, D., Xu, J., D'Haro, L., Haro, D., & Ney, H. (2006). Error analysis of statistical machine translation output. *Proceedings of the Fifth International Conference on Language Resources and Evaluation,* 697-702.

Appendix A

Online Links of Commencement Speeches Given by Male Speakers

This table presents the list of male commencement speakers, together with the links to the videos in which their speeches appear. The transcript to each speech is also linked below. These transcripts are taken from various sources. When the official source could not be accessed, or was not available, third-party sources, such as singjupost.com and speakola.com, were utilized. The first link is the recorded video and the second is the transcript.

N	Speaker	Links of the Commencement Speech on You Tube & Utilized Transcripts for the Research
1	Barack Obama	https://www.youtube.com/watch?v=Iz3Z3jNDvQ4 https://www.politico.com/story/2016/05/obamas-howard-commencement-transcript-222931
2	Bill Gates	https://www.youtube.com/watch?v=zPx5N6Lh3sw&t=241s https://news.harvard.edu/gazette/story/2007/06/remarks-of-bill-gates-harvard-commencement-2007/
3	Conan O'Brien	https://www.youtube.com/watch?v=ELC_e2QBQMk https://www.dartmouth.edu/~commence/news/speeches/2011/obrien-speech.html
4	Eric Schmidt	https://www.youtube.com/watch?v=GZRIMQxje7k https://singjupost.com/transcript-googles-eric-schmidt-at-2012-boston-university-commencement-speech/
5	Jeff Benzos	https://www.youtube.com/watch?v=zbRQ2Dlqde8 https://www.princeton.edu/news/2010/05/30/2010-baccalaureate-remarks
6	Mark Zuckerberg	https://www.youtube.com/watch?v=BmYv8XGl-YU&t=603s https://singjupost.com/facebook-ceo-mark-zuckerbergs-harvard-commencement-speech-full-transcript/?singlepage=1
7	Robert De Niro	https://www.youtube.com/watch?v=7Y1hyAf8xy0 https://speakola.com/grad/robert-de-niro-next-tisch-201
8	Steve Jobs	https://www.youtube.com/watch?v=UF8uR6Z6KLc https://singjupost.com/full-transcript-steve-jobs-stay-hungry-stay-foolish-speech-at-stanford-2005/
9	William McRaven	https://www.youtube.com/watch?v=pxBQLFLei70 https://news.utexas.edu/2014/05/16/mcraven-urges-graduates-to-find-courage-to-change-the-world/
10	Tim Cook	https://www.youtube.com/watch?v=2C2VJwGBRRw&t=161s https://news.stanford.edu/2019/06/16/remarks-tim-cook-2019-stanford-commencement/

Appendix B

Online Links of Commencement Speeches Given by Female Speakers

This table presents the list of female commencement speakers, together with the links to the videos in which their speeches appear. Just like with the list of speeches given by males, the transcript to each speech is also linked below. These transcripts are also taken from various sources, both official, in this case the university website, and third-party, such as singjupost.com and speakola.com. The first link is the recorded video and the second is the transcript.

N	Speakers	Links of the Commencement Speech on You Tube & Utilized Transcripts for the Research
1	Hillary Rodham Clinton	https://www.youtube.com/watch?v=YJFAByAtC4U&t=1028s https://speakola.com/grad/tag/HILLARY+CLINTON
2	Joan K. Rowling	https://www.youtube.com/watch?v=wHGqp8lz36c&t=601s https://singjupost.com/j-k-rowling-speaks-harvard-commencement-2008-transcript/
3	Mary Barra	https://www.youtube.com/watch?v=6FCZWPzmnNc https://www.mlive.com/news/ann-arbor/2014/05/read.html
4	Meryl Streep	https://www.youtube.com/watch?v=5-a8QXUAe2gcom/watch?v=5-a8QXUAe2g https://mlive.com/meryl-streep-barnard-commencement-speaker-2010-columbia-university-full-transcript/
5	Michelle Obama	https://www.youtube.com/watch?v=MgqAhn0a-tk https://www.ccny.cuny.edu/commencement/commencement-address-first-lady-michelle-obama
6	Mindy Kaling	https://www.youtube.com/watch?v=JgUDjixWB5I https://news.dartmouth.edu/news/2018/06/2018-cmmencement-address-mindy-kaling-01
7	Natalie Portman	https://www.youtube.com/watch?v=jDaZu_KEMCY&t=210s https://singjupost.com/full-transcript-natalie-portman-harvard-commencement-speech-2015/
8	Oprah Winfrey	https://www.youtube.com/watch?v=GMWFieBGR7c https://singjupost.com/oprah-winfrey-speech-at-harvard-commencement-2013-full-transcript/
9	Savannah Guthrie	https://www.youtube.com/watch?v=_NuEVg8u98g https://www.today.com/news/read-savannah-s-guthries-full-graduation-speech-gw-s-class-t154411
10	Sheryl Sandberg	https://www.youtube.com/watch?v=8w1d1TWxwec http://news.mit.edu/2018/sheryl-sandberg-commencement-address-0608