Evaluating Automated Speech Recognition Captioning and Its Possibilities of Language Learning

Alvin Taufik[1], Tiffany Tan[2]
[1,2]Department of English Language and Culture
University of Bunda Mulia, Jakarta, Indonesia

Abstract: *Ask everybody about the quality of machine translation in 2017, and it is almost certain that most will say that it hasn't been satisfactory yet. Such is also the case for automatic captioning (AC) that can be found on video based website such as Youtube, or is it? This research is aimed to investigate the quality of ASR found on Youtube channels and videos. We have also learned a lot about the use of this in language learning. In this research, the additional purpose is to find out whether the investigation on the quality of the caption can be used to assist in language learning. Additional objective of this research is to familiarize the students with new technology. This is a descriptive qualitative research. The research is conducted through observations of Youtube channels and videos about different interests. The research will identify the problem related with the ASR, and its possible solution in relation to language learning.*
Keywords: *Automatic captioning, language learning*

I. INTRODUCTION

1.1 Background

How many of you have never used google translate? The truthful answer for such question is probably: none. From professionals who use google translate for their work, or those who use it simply because they want to find the equivalence of some words in their source language, thousands are using it everyday. From the many users, there is a common unwritten consent that although the results of machine translation hasn't been satisfactory, they are improving.

So, can it be said that all those efforts in creating machine translation (MT) is not yet successful? Actually, in some cases, it has been proven to be successful. Hutchins (1999) has created lists of successful MT systems. One of them is Météo, a translation system for Canadian weather reports. The language pair is English and French. This one is probably the most quoted MT system in practical use (quoted in Volk, 2008).

Other 'success' stories regarding MT include the usage of METAL in specific banking and trade industry, and customized Systrans. However, as it is today with Google Translate, they required post editing. There was one project which limits the post-editing phase, yet this project controls the language being transfered (Hutchins, 1999).

So, how about automated or real-time caption which are generated by speech recogniyion then? Unlike general ideas of MT, automated speech recognition, or from now on refered to as ASR is said to have an advantage over conventional translation in regards with MT (Popowich, et. al., 2000). The first advantage is, subtitle needs to be presented with a correct grammar. This accomodates the rule based approach of ASR. There is also an ASR based on example yet

nothing conclusive can be taken from such approach. The same can also be said for statistical based approach for MT on film. Furthermore, eventhough ASR is benefited by the use of proper grammar, the observation has proven that it still has disadvantages. A quick look at the system has shown that it is still struggling with peoper names and nouns. Another observable mistakes is the caption is created based on the sounds produced by the speaker in the video. A video which has been proven to be unclear in its pronunciation has been observed to also have subpar captions.

This disadvantages that it still has can be used as an alternative to language learning. One use of the ASR that the researcher can think of at the moment is error observation and correction. There are many other forms of lesson which can be made based on this concept. A further research will show what else can be done to ASR in language learning.

1.2 Statement of Problem and Research Question

To successfully deal with real-time captioning, and its possible manipulation for language learning, there are several problems which needs to be addressed firsthand. As stated above, a quick observation has shown what are still wrong with ASR. However, has it ever been proven to have more problem than that? Or will that be the only problems with ASR. A second problem is related with the format of the ASR itself. Before addressing the problem, one must know how it works. In that case, the researcher must know the concept of ASR. The third problem is with its use for language learning. A researcher must questions the best way to integrate the ASR into classroom situation.

II. THEORETICAL REVIEW

Automated Speech Recognition (ASR)

This kind of research is also known in another name; a speech-to-text synchronization problem, or as an imperfect transcript correction problem. They mainly deal with the low quality of provided transcripts or captions. One website which has been using ASR is Youtube. This largest video channels in the world shares million of videos everyday on different fields of study and interest. Only recently these videos have been available for people who needs ASR the most, the deaf and hearing impaired (DH). However, challenges still remains in providing the DH, or public audience in general, with proper ASR.

So, how does this ASR work? Youtube (which was bought by Google) has been using caption since 2008, and in 2009 they decided to go with machine-generated caption using speech-recognition model. This model has acoustic, lexicon, and language components (Greenemeire, 2011). The acoustic model is a statistic-based model. The lexicon consists of list of words and their pronunciation, and pronunciation variations.The language component is also a statistical model of phrase and sentence which is used in a language.

One of the problems in creating an excellent ASR is the separation of the source. Many shared videos are poor in quality. This makes it difficult for the machine to separate between language spoken and background noises. It also has problems in context-specific words, such as the one mentioned in a lecture or some sorts. Another problem with ASR, according to Panayota (2012), is accent. In countries where English is spoken as second language, the English spoken by the locals are often misinterpreted because the accent of said people is still apparent. This

leads to a lot of editing from the native speaker of English. It is interesting to find out whether accent is still an issue in ASR today.

Some research on this has used models of distance in determining the equivalents. The first one is Hamming distance model. In summary this model measures 'the minimum number of substitutions required to change one string into the other, or the minimum number of errors that could have transformed one string into the other'. Using this distance unit, we can find out the number of substitution or errors have been made when changing one string (in this case is the spoken language) into another (caption). Another measurement units which have been used to evaluate MT product is the Levenshtein distance. This units includes the calculation of deletions, insertions, and substitutions.

These deletions, insertions, and substitutions rate are later converted into Word Errors Rate (WER). The example of these WER is as follows:

> Let us say that the audio reads:
>
> 'we have never included his name in it'
>
> And the transcription is given:
>
> 'we haven't never include it his name in it'

From this example, there are two insertions in 'n't' and 'it' and one deletion in 'd' in 'included'. Therefore, the WER is 3/8 or 37.5%.

III. RESEARCH METHOD

3.1 Research Design

This research is a descriptive qualitative research. The research is done through observations of Youtube channels and videos on different interests. The research will identify the problem related with the ASR, and it also seeks the possible solution for the problem. Additionally, it also investigates the possibilities of its application in language learning.

3.2 Data Source

The data to be used in this research is Youtube videos which have been fitted with automatic captioning. Per data acquired in 2011, there have been 60 million videos which have been auto-captioned. For this research, however, the researcher will choose one video for each subject. The one video which will be chosen is chosen based on the poster. Youtube videos can be separated into three based on its poster. There are official videos made by the official channel. There are 'special' channel that post videos taken from the official channel. Most of the time, the video from both channel will be the same. Finally, there are videos posted by individual. This video is usually of poor qualities. On this research, only videos from 'special' and official channels will be used. The reason why videos from individual channels are not chosen is because they are usually very poor in quality, and because of that, they present the least interest to the audience. One other limitation to the data is on the format of the videos, which will be speech-heavy. This research will be a descriptive-qualitative research because the data to be used is in

the form of errors found, and not the quantification of the errors. This research will describe the errors found, and its possible solution. Furthermore, this research will use th data as a source of teaching.

3.3 Data Analysis Procedure

The data will be observed for possible errors. Having found the errors, the data will later be codified for convenience. After the codification, the research will analyse the data using the theories provided in the theoretical framework. After the data analysis is completed, the data will later be investigated for a possible solution. The solution will later be included as alternatives when the video is used in the language learning.

3.4 Implication

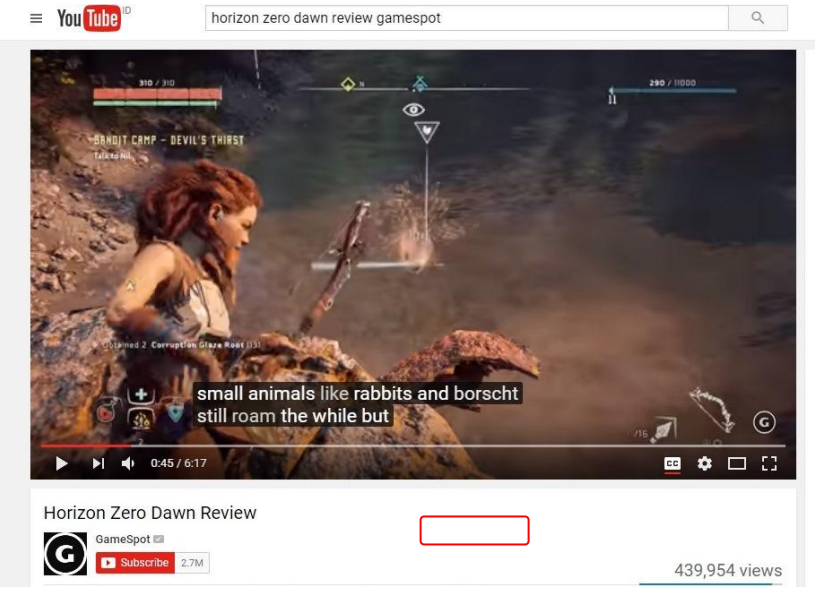
The outcome of this study is expected to provide an in-depth description of the errors found in the use of MT in ASR. In addition, the videos used in this research, and the research finding can be integrated into language learning and teaching. Further benefits will be the use of technology in language classroom, aand as alternatives in teaching language.

IV. DATA ANALYSIS, RESULTS, AND DISCUSSION

4.1 Results and Analysis

As mentioned in the From the observation done on the Youtube videos and their automated caption, here are the results and analyses. The first is the result from a channel called 'Gamespot'. This channel discusses videogames, so it will be heavy with videogames jargon and proper names. Take a look at the example below:

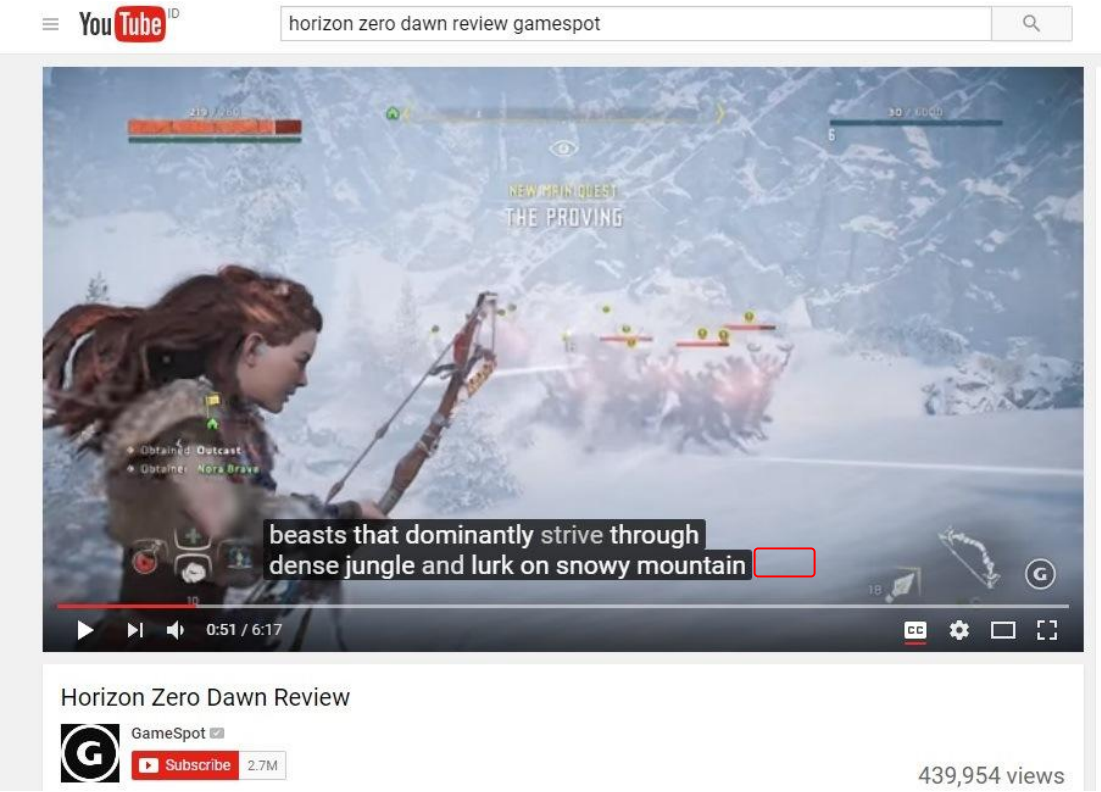
Figure 5.1 Videogame Errors: Wrong Words



The autogenerated caption (AGC) in this video transcribe both the in-game dialogues and the narrated review of the videogame being reviewed. Early observation shows that the in-game dialogues are not transcribed as complete as the narrated review. As an example, when one of the characters said 'keep pressing', the AGC was only able to transcribe 'keep'. The narrated review, on the other hand, has generated a decent transcription. The writer noticed two errors when the review discussed a name of an animal and what they do. The narrated review actually said, 'small animals like rabbit and boar still roam the wild...', however, the AGC failed to caption the second animal name and the last word in that sentence. The caption becomes 'small animals like rabbit and borscht still roam the while...',

Further observation of the AGC shows that other than mistranscribing the ending of a sentence, it might also eliminate them completely. Take a look at the illustration below:

Figure 5.2 Videogame Errors: Missing Ending



In this scene, the original ending of the sentence is actully read '....and lurk on snowy mountain tops'. However, as can be seen in the illustration, the word 'top' has been eliminated in the transcription. Another error found in the AGC is a problem in the pronunciation. Look at illustration 5.3 for this example.

Figure 5.3 Video game Errors: Mispronunciation of Ordinary Words

From this illustration, it is seen that the word ‘combat’ has been written in ‘kombat’. Another interesting pronunciation error can be seen in the names of the characters. Have a look at illustration 5.4.

Figure 5.4 Videogame Errors: Mispronunciation of Proper Names

From the side-by-side comparison, we can see that the AGC cannot get the pronunciation of the name clearly in both ocassions. The main character's name in this game is 'Aloy'; on the left screen, as can be seen, although it gets the letters correctly, it makes a mistake in capitalization, as the word 'Loy' is capitalized, as if it is a noun or other parts of speech. On the right screen, it missed the transcription completely by using the letter 'r' instead of 'l' in it, and giving capitalization in the beginning of the word 'Roy'.

The error rate for the video is counted based on the sentence. The sentences included in the calculation are based on both in-game dialogue and narrated review. The results of the calculation is presented in the table below.

Table 5.1 WER of AGC in Videogame Context: 1st Results

No.	Sentence Origin	WER (in average)
1	In-game Dialogue	44.8%
2	Narrated Review	5.98%

As seen from the table above, the in-game dialogue transcription produces a very high error rate. This is with the audio produced in the same quality as the narrated review. An amazing result is seen on the narrated review AGC. From 55 sentences being transcribed, 27 of them have been done flawlessly by the machine. The ones having erronous transcription have an average WER rate below 20%; resulting in a very high accuracy of 94.02%.

The researcher has conducted observation of other videos with the same genre. The observation is conducted to solidify the reliability of the AGC. The finding of the error rate of these can be seen below.

Table 5.2 WER of AGC in Videogame Context: 2nd Results

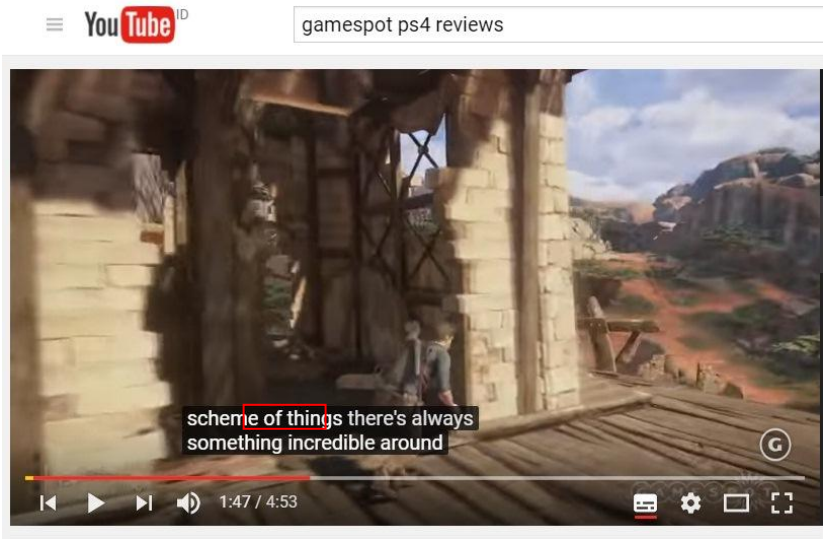
No.	Sentence Origin	WER (in average)
1	In-game Dialogue	27.6%
2	Narrated Review	0.51%

In addition to the rsults of WER, a new finding has been found on the second testing.

Figure 5.5 Videogame Errors: Missing Apostrophes

From figure 5.5., we can see that the apostrophe which is supposed to exist in the phrase is missing. This occurence is actually quite prominent in the analysis. This is considered as an error because it is not the fault of the system. Unlike the obvious missing punctuation, which is probably the fault of the system since it is not found in any videos observed by the researcher, the apostrophes can still be seen in other part of the video, such as seen below.

Figure 5.6 Videogame Errors: Missing Apostrophes – Comparison



To further validate the results, another testing is conducted on the WER of the same genre video. The result of the third testing can be seen below.

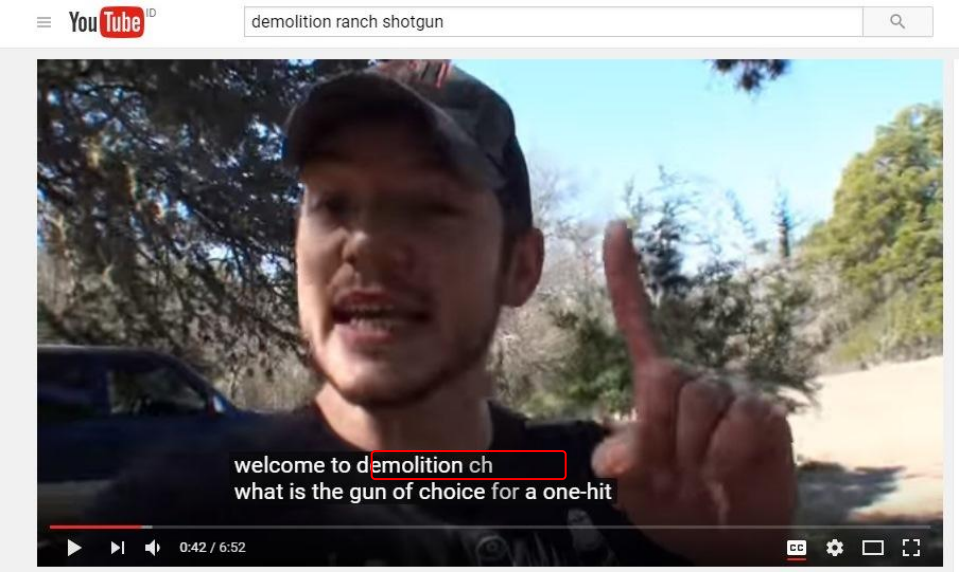
Table 5.3 WER of AGC in Videogame Context: 3rd Results

No.	Sentence Origin	WER (in average)
1	Narrated Review	3.73%

The results shown by the additional two tests are consistent with the result acquired from the first test. It has an average of 97.88% accuracy rate, or 2.12% WER. The 3rd video actually contains a lot of mistakes in it, yet they are limited to phrasial errors.

The second WER analysis is from a channel called 'Demolition Ranch'. This channel has videos about guns, so it will be filled with gun terminologies. The first mistake observed from this channel is on the incomplete end-sentence word, which happens to also be the name of the channel. Take a look at the illustration below.

Figure 5.7 Gun Terminologies Errors: Incomplete Proper Names



This phenomenon might be explainable by the speed of which the sentence is spoken. Unlike the example found in illustration 5.2 the word 'ranch' in the phrase 'demolition ranch' which is supposed to end the sentence is spoken very rapidly. So it can be assumed then that the speed of the speech defines the AGC as well.

A very interesting error was made by the ACG on another proper names. See illustration below:

Figure 5.8 Gun Terminologies Errors: Wrong Word

As can be seen from it, the word 'magnum' has been replaced by 'magnet'. This is similar to the example found on the illustration 5.1. One difference is, in illustration 5.1., the word is not recurring. This one, on the other hand, does. The same phrase was stated earlier in the video, as seen in illustration 5.7., yet, the ACG transcribed it accurately. If seen from the speed of the speech, they are uttered in a similar pace. In terms of the word which collocates it, they are also similar. The reason for the error is, then, unknown. Further study is needed to investigate this.

Figure 5.9 Gun Terminologies Errors: Comparison

The error rate for this video is even lower. From more or less 62 sentences, only three are aerroneous. The errors have been described above. Here are the the WER table for gun terminologies video:

Table 5.4 WER of AGC in Gun Context

No.	Sentence Origin	WER (in average)
1	Narrated Explanation	1.14%

As can be seen from table 5.2., a very good result is shown from by the AGC of the narrated explanation. From 62 sentences being transcribed, only 3 (three) of them have errors. This leads to a very high accuracy rate of of 98.86%. The similar results are also visible from the second testing, as shown in table 5.5. below, in which produced a 97.6% accuracy rate.

Table 5.5 WER of AGC in Gun Context

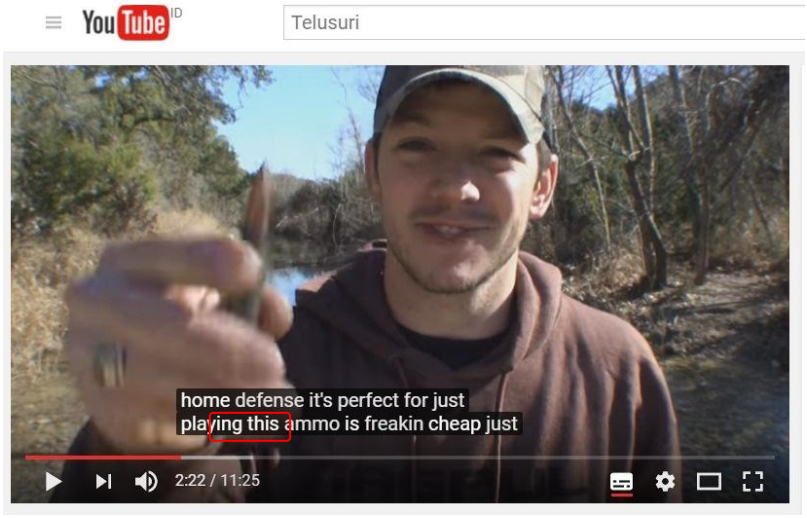
No.	Sentence Origin	WER (in average)
1	Narrated Explanation	2.40%

As in the video game context, after the second testion, there is an additional finding. Here is an example.

Figure 5.10 Gun Terminologies Errors: Improper linking verbs

As can be seen from the illustration, after the proper name ‘Mosin Nagant’, there should be a linking verb ‘are’. Yet, it has been replaced by an unknown, perhaps meaningless ‘SAR’. Furthermore, this occurence is repeated in all genre of videos being analyzed. In addition, unlike the inexistence of punctuations, which is probably the system’s limitation, the linking verbs are actually transcribed in other part of the video as can be seen from figure 5.11 below.

Figure 5.11 Gun Terminologies Errors: Improper linking verbs



The third testing is rather different than the first two. In the first two tests, the speakers are native speakers of English or people who have embraced their native speaker intuition and spoke in somewhat native speaker's accent. In this third one, however, the researcher observes video in which the speakers are still influenced by their local accent. The first video contains accent-heavy Indian speaking in English.

From the observation, it was found that the errors are similar to those found in native speakers' video. Proper names are still the most common errors found in the video (see figure 5.12). One 'anomaly' found in accented videos are the repeated words. Unlike the first two tests; one being a video review in which the script has probably been prepared and practiced, and the other being a monologue, this one is in a form of dialogue. Perhaps this is the reason why the repeated words in the video are transcribed (see figure 5.13)

Figure 5.12 Accented Videos' Errors: Inappropriate Proper Names

Figure 5.13 Accented Videos' Errors: Repeated Words

It turns out that the WER of the accented videos are not too dissimilar with the one spoken by native speaker with an average of 0 – 1 % WER or more than 95% accuracy rate. In conclusion, from the seven analysis on the WER, it can be said that the accuracy rate of the AGC is around 95%. This is consistent with previous studies. Moreover, proper names have become the most convoluted area of the errors.

However, it is safe to say that not all types of speech is covered in this research. From the initial observation done on videos with shouting involved, it seems that the error counts can be more apparent. It is interesting to know the WER results of such video. In addition to that, perhaps one of the most glaring weakness of this AGC is the lack of punctuation or capitalization. While this might not be a problem for those who can hear perfectly, for the hard-of-hearings and, especially, the deaf, this might be a huge problem. If the AGC is intended to provide accessible media for all, then this is the first problem which needs to be fixed.

4.2 Discussions: Teaching / Learning Applications

The idea of using caption to teach or learn english has been around since the 1990s; some say it's even earlier than that. However, the captions which were found in the films or even television has been edited before it went air. As a result, there is very limited applications for the captions.

Automatic Generated Caption, however, tells a different story. One major difference is, they are not edited. Non-edited captions mean that there are going to be errors. These errors mean that the option to use them as a teaching or learning tools, are not as limited as its predecessors. Another difference is in its form. TV's or film's captions, regardless of the

audience understanding of it, are often a reduced versions of what is actually spoken in the screen. These too can be used to enrich a teaching or learning experience.

In this paper, the researcher is focusing his attention on using AGC to enrich the learning experience of a business English student. There are several reasons for the choosing of the class. First, this class is an English department class. It is hoped that English students will have a more profound understanding of the captions which are presented in English. Second, the students of this class are now in their fifth semester. Having learn basic English skills in the first until fourth semester, the researcher expects the students to be able to analyze the various types of errors which are found in Youtube videos. The application of the AGC to the teaching and learning of Business English will be focused on its material development.

In developing the materials, there are several aspects which need to be considered, they are contextualization, resources, individual needs, and timeliness (Howard and Major, 2004). First, the developer needs to pay attention to context. It is known that most reference books are anglocentric. As a result, they are heavily influenced by the context which happens in the EFL countries, but might not happen in the local setting/s. The second is about resources. Each teaching and learning activity has its own condition. Some institutions will have the most complete resources such as the access to internet, full set of audio, or even visual materials, while some rely heavily on chalk and board. The students' resources are also a critical point in developing materials. We cannot make materials which cannot be fully utilized by the students. In other words, in developing materials and their application, teaching and learning condition must be considered.

Another aspect worth mentioning is the individual needs. With the internet today, any individual can be connected to the web. With this in mind, accessibility is no longer an issue. In relation to this research, when accessibility to the internet is open to everybody, every student can select their own materials which will suit individual needs. In designing their own materials, teachers can also develop courses which are up-to-date. Coursebooks might include recent events. However, by designing materials teachers can acutally make use of current events in their teaching so that it can be more relatable with the students. Youtube is updated daily, so, it is possible to have recent and relatable events which can be integrated into their lessons.

It can be summed up then that in developing materials, teachers need to consider their learners, context, and resources. To know about the learners, material designers need to do some needs analysis to find out about learners'needs in relation to language skills, learners' preferences, learners' experiences and knowledge of their first language and the language they are trying to learn, learners' interest and purposes in learning the language. Context refers to timeliness and socio-cultural appropriateness. Timeliness refers to the relatability of the materials with the students; usually refers to current events. Socio-cultural appropriacy is often related to the 'real' situation that the learners might face. This means that the materials designed must adequately represents the real language condition in which the language will be used.

Finally, teachers must be aware of their resources. This is not only about the facilities in said institution, but also on the leaners' learning condition. This leads to the inclusion of the resources in the needs analysis. Internet is a great way to balance this situation since access to it has become easier and cheaper. To answer the challenges stated above, the researcher utilize Youtube channels since they confirm to the area of limited resources, timeliness, and contextual elements.

V. CONCLUSION

From the analysis and discussion above, the answers to the research questions have been acquired. First of all, to the question on the problems with AGC or ASR from Youtube, it is found that the system still has problem in transcribing the proper names, phrasing problems, and fast pace speech which resulted in incomplete words or sentence endings. These problems also show how the AGC system works. As stated earlier in the Review, the AGC works in three phases; acoustic, lexicon, and language component. As acoustic is related to sound recognition, the problem with proper names might have started from this. This is obviously caused by the very large probabilities and variations of proper names pronunciation. This problem might not be solvable if the system is still statistical since it is clearly impossible to input all proper names and its pronunciation variation into the system.

In phrasing problem, this might happen in the third stage: language component. In this stage, the system deals with commonly used phrases and sentences. Due to the lack of input of some phrases found in the video, or some mistakes in the binary matching, a wrong word is used. This problem can be solved if the new phrases are included into the system as a part of editing process by either the writers or the audiences. The third problem is also a classic human error. If the speech is uttered to quickly, logically, even human cannot get all the utterance, let alone the machine. This is very solvable, especially if the poster of the videos is posting them for the purpose of education. The third research questions cannot be answered conclusively at the moment since testing is still needed. However, in the teaching, there are some factors which might have to be considered by the educators. They are context, timeliness, individual needs, and resources

REFERENCES

[1] Georgakopoulou, P. (2012). Challenges for the audiovisual industry in the digital age: the ever-changing needs of subtitle production. *The Journal of Specialised Translation 17: 78-103*.

[2] Greenemeier, L. (2011). Say what? Google works to improve YouTube autocaptions for the deaf. *Scientific American*.

[3] Howard, J., and Major, J. (2004). Guidelines for designing effective English language teaching materials. *The TESOLANZ Journal* 12: 50-58.

[4] Huang, C, et al. (2003). Automatic closed caption alignment based on speech recognition transcripts. *Rapport technique*, Columbia.

[5] Lecouteux, B, et al. (2006). *Imperfect transcript driven speech recognition*. InterSpeech.

[6] Panayota, G. (2012). Challenges for the audiovisual industry in the digital age: the ever-changing needs of subtitle production. *Journal of Specialized Translation*. Issue 17. pp.(78-103). Retrieved from http://www.jostrans.org/issue17/art_georgakopoulou.pdf.

[7] Volk, M. (2009). The automatic translation of film subtitles. A machine translation success story?. *JLCL 24.3: 115-128*.

[8] Yim, J. (2016). Design of a Subtitle System for Internet TV Systems. *International Journal of Multimedia and Ubiquitous Engineering 11.5: 11-20*.