CHAPTER I

INTRODUCTION

1.1 Background of the Problem

The quest for a better understanding in learning languages has been great concern for both learners as well as educators. One of indicators of successful language learning is shown by the abilitiy in producing words as the way native speakers do or sequencing words side by side correctly (Setiarini, 2018). This sequencing and co-occuring words are what Firth (1957) introduced as collocation. Firth defines collocation as the relationship between a word and its surrounding context where frequent co-occurence with other words or structures helps to define the meaning of the word.

Collocation is classified into two main categories. They are lexical collocation and grammatical collocation. This distinction was first made by Benson and Ilson (1986). BBI, The Combinatory Dictionary of English, defines a grammatical collocation as a phrase consisting of a dominant word (noun, adjective,verb) and a preposition or grammatical structure such as an infinitive or clause, while lexical collocation as a phrase consisting of nouns, adjectives, verbs and adverbs. The basic difference between grammatical collocation and lexical collocation is that the former contains word together with preposition, infinitives or clause, while the latter does not.

English learners might still have difficulties in understanding and using English grammatical rules especially collocation. The way they apply the rules is influenced by their first language. According to Mahmoud (2005), EFL learners depend on interlingual strategies to facilitate learning. The phrase “same as” and

“same with” are examples which english learners tend to rely on their mother tongue when producing english utterance. Language leaners might fail to translate this collocation correctly when they translate “dengan” into “with” instead of “as”, which in Indonesian both mean the same thing basically. In case of native English speaker, the word “with” does not frequently co-occur after the word “same”, thus it might sound unnatural to use it along with the word “same”

The collocation of source language text (SLT) cannot be translated literally into target language text (TLT). This word-by-word tendency is what might later hinder language learners’ fluency. An example that also has similar problem to this research topic is the phrase “different with” instead of “different from” since in Indonesian the phrase means “berbeda dengan”. According to Sari and Gulo (2019) this shows that the way the learners use English is influenced by their mother tongue or primary language. Because in Bahasa Indonesia the word “with” means “dengan”, the learners tend to produce the phrase “different with” in their speaking or writing. This phenomenon likely to happen in various English utterances produced by language learners for example the phrase “terima kasih sebelumnya” becomes “thanks before” instead of thanks in advance, “berbeda dengan” translated into “different with” rather than different from, ”mirip dengan” becomes “ similar with instead of “similar to, “bekerja dengan” becomes “work with” instead of work for, “tertarik dengan” as in “interested with” contrary to “interested in”, and others which share the same root problem.

Langugage fossilization is what might later occur if English learners keep repeating the same mistakes/errors continously and unconsciously. According to Selinker (1972) language fossilization refers to certain aspects which speakers of a particular native language (L1) will tend to keep in their interlanguage relative to a

particular target language (L2) no matter what the age of the learner or the amount of explanations he receives in the target language. From the definition above, language fossilization can also be inferred as when wrong input produces wrong output by language learners which later remained immovable. Thus, it is needed to point out this phenomenon in order for language learners to reach a better understanding and hopefully improve their target language, in this case English, specifically the aspect of grammatical collocations.

Greenbaum (1974) argued that some collocations can be identified intuitively particularly for obvious cases of collocation. However he also tried to address that intuition can be typically a poor guide to collocation. He found that people disagree on collocations. This is also supported by Krishnamurthy (2000) that each of us has only a partial knowledge of the language, we have prejudices and preferences, our memory is weak, our imagination is powerful, and we tend to notice unusual words or structures but often oveerlook ordinary ones. thus it may not be reliable to have intuition as measurement.

There is no total agreement among native speakers as to which collocation are acceptable and which are not (Partington : 1998). While it is possible to observe collocation informally using intuition, Hunston (2002) pointed out that it is more reliable to measure collocation statistically, and for this a corpus is essential. Klimov (2013) stated that corpus can draw a large amount of authenthic, naturally occurring language data by speakers or writers in order to confirm or refute researchers' hypotheses about specific language features based on empirical data. Therefore researchers do not have to rely on their own or other native speakers' intuition or even

some made-up examples, rather using corpus as a better alternative to measure and reveal the truth of the words.

Due to the complexity of English collocation, this paper is going to focus on grammatical collocations “same as” and “same with” especially found in the computerized database for linguistic research COCA (The Corpus of Contemporary. American English) in various genres which corpus COCA provided with. Based on the consideration above, the researcher is interested in examining this phenomenon and propose a research title : **A Corpus-Based Analysis of Grammatical Collocations “Same As” and “Same With”**

1.2 Formulation of the Problem

Research Questions :

1. How are grammatical collocations “same as” and “same with” compared in terms of the frequency based on the data found in the corpus COCA ?

2. What are the commonly possible context or function which the phrase “same with” might happen as a better alternatives than the phrase “same as”.

1.3 Purpose of the Research

1. To describe the comparison of grammatical collocations “same as” and “same with” in terms of the frequency based on the data found in the corpus COCA.

2. To know what common context or function could the phrase “same with” possibly be used instead of “same as”.

1.4 Significance of the Research

1. The researcher hopes to promote the utility of corpus linguistics especially the corpus of contemporary American English (COCA) in examining and uncovering linguistic features based on actual usage for language teaching and learning.

2. This study hopes to help raising awareness to English learner on the importance of grammatical collocations and bring benefits to those who are interested in corpus study.

1.5 Limitations of the Research

This research is going to focus on the usage of grammatical collocations of "same as" and "same with," particularly in determining their frequency and commonly possible context or function using COCA (the corpus of contemporary American English). To help the researcher narrow the extensive data, the research object will be based on large collection of "real life" language usage stored in the corpus (computerized database for linguistic research) using purposive sampling technique which enables the researcher to select certain data that suits the criteria of the study. The data will be based on various genres which COCA provided with in order to ensure the credibility of this study findings and conclusion.

The corpus of contemporary American English (COCA) is widely-used and genre-balanced. The corpus contains more than one billion words of text

approximately 25+ million words each year from 1990 until 2019 with eight different genres. The genres consist of spoken, fiction, popular magazines, newspaper, academic texts, TV and movies subtitles, blogs, and other web pages. All eight genres consist of different sub-genre to narrow down and specify the data. Moreover, COCA also provides various data based on the years. This means it enables users to map out recent changes in English.

Genre	# texts	# words	Explanation
Spoken	44,803	127,396,932	Transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), Oprah)
Fiction	25,992	119,505,305	Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, and fan fiction.
Magazines	86,292	127,352,030	Nearly 100 different magazines, with a good mix between specific domains like news, health, home and gardening, women, financial, religion, sports, etc.
Newspapers	90,243	122,958,016	Newspapers from across the US, including: USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle, etc. Good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc.
Academic	26,137	120,988,361	More than 200 different peer-reviewed journals. These cover the full range of academic disciplines, with a good balance among education, social sciences, history, humanities, law, medicine, philosophy/religion, science/technology, and business
Web (Genl)	88,989	129,899,427	Classified into the web genres of academic, argument, fiction, info, instruction, legal, news, personal, promotion, review web pages (by Serge Sharoff). Taken from the US portion of the GloWbE corpus.
Web (Blog)	98,748	125,496,216	Texts that were classified by Google as being blogs. Further classified into the web genres of academic, argument, fiction, info, instruction, legal, news, personal, promotion, review web pages. Taken from the US portion of the GloWbE corpus.
TV/Movies	23,975	129,293,467	Subtitles from OpenSubtitles.org, and later the TV and Movies corpora. Studies have shown that the language from these shows and movies is even more colloquial / core than the data in actual "spoken corpora".
	485,179	1,002,889,754	