The Graduate Institute of Sciences and Engineering

M.Sc. Thesis in Electrical and Computer Engineering

EVALUATION OF ONLINE TRANSLATION SERVICES' OUTPUT QUALITY

by

Mehmet Akif GEDIK

February 2015

Kayseri, Turkey

EVALUATION OF ONLINE TRANSLATION SERVICES' OUTPUT QUALITY

by

Mehmet Akif GEDIK

A thesis submitted to

the Graduate Institute of Sciences and Engineering

of

Meliksah University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Electrical and Computer Engineering

February 2015
Kayseri, TURKEY

APPROVAL PAGE

This is to certify that I have read the thesis entitled "Evaluation of Online Translation Services' Output Quality" by Mehmet Akif GEDIK and that in my opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science in Electrical and Computer Engineering, the Graduate Institute of Science and Engineering, Meliksah University.

_____

February 25, 2015 Assist. Prof. Hasan KITAPCI
Supervisor

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

_____

February 25, 2015 Prof. Dr. Murat UZAM
Head of Department

Examining Committee Members

Title and Name		Approved
Assist. Prof. Dr. Hasan KITAPCI	February 25, 2015	_____
Assoc. Prof. Dr. Ahmet UYAR	February 25, 2015	_____
Assist. Prof. Dr. Mete CELIK	February 25, 2015	_____

It is approved that this thesis has been written in compliance with the formatting rules laid down by the Graduate Institute of Science and Engineering.

_____

Assist. Prof. Dr. M. Evren SOYLU
Director

February 2015

EVALUATION OF ONLINE TRANSLATION SERVICES' OUTPUT QUALITY

Mehmet Akif GEDIK

M.S. Thesis – Electrical and Computer Engineering
February 2015

Supervisor: Assist. Prof. Dr. Hasan KITAPCI

ABSTRACT

In many environments such as social, multimedia, education, news, politics, etc. people need to learn more about new information from foreign sources more than past. Since everybody cannot know any foreign language as well as an expert, they need to obtain meaning of texts from computers with correct translation of texts immediately or by human support manually. Every text cannot be translated by human labor immediately and fast. Preparing correct translation for every text is hard, cost is high and it takes so long time for experts. Computers provide candidate translations but their correctness levels are unknown.

In this research, a comparative evaluation about output quality of online machine translation services was performed on a dataset collected from a randomly selected bilingual sentence pairs in English and Turkish languages. Some sentences are used for training and others are used for verification. Sentences are categorized based on their structure types and statistical analysis on word counts done for better evaluation results, coming from 4 different essential bilingual corpora, which contain source and human reference translation sentence. They are compared with sentences coming from popular online translation services **Google, Bing and Yandex** using some most popular and successful evaluation methods such as **Precision**, **Recall**, **Bleu, Meteor** and **Bleu+**, which is an eligible approach for agglutinative languages like Turkish. Then, human evaluation comparison tests were done to compare the human approach and automatic evaluation results to measure output quality of online machine translation services better correlated with expert judgment.

Keywords: Output Quality of Online Machine Translation Service, Evaluation of Machine Translation, Precision, Recall, Bleu, Meteor, Bleu+

ÇEVRİMİÇİ DİL ÇEVİRİ SERVİSLERİNİN ÇIKTI KALİTESİNİN DEĞERLENDİRİLMESİ

Mehmet Akif GEDİK

Yüksek Lisans Tezi – Elektrik ve Bilgisayar Mühendisliği
Şubat 2015

Tez Yöneticisi: Yrd. Doç. Dr. Hasan KİTAPÇİ

ÖZ

Sosyal, çoklu ortam ve eğitim gibi pek çok sahada insanlar yabancı kaynaklardaki yeni bilgilere her zamankinden daha fazla öğrenme ihtiyacı duyuyorlar. Her insan yabancı bir dile bir uzman kadar hâkim olamayabilir. Bu durum insanlara metinlerin doğru tercümelerinin otomatik olarak ya da insan eliyle sağlanmasını gerektiriyor. Her metin insan eliyle ani ve hızlı bir şekilde çevrilemez. Her metne doğru tercüme hazırlamak uzmanlar için zor, masraflı ve uzun zaman alan bir iş yüküdür. Makineler ise bize yaklaşık bir tercüme vermektedirler fakat doğruluk seviyeleri bilinmemektedir.

Bu araştırmada, makine tercüme servislerinin çıktı kalitesinin değerlendirme karşılaştırması, kaynak cümle ve insan referans tercümesine içere iki dilli Türkçe ve İngilizce cümle çiftleri üzerinde uygulanmıştır. Bu cümle çiftlerinin bazıları eğitim ve diğer kalan kısmı teyit testleri için kullanılmıştır. Daha iyi bir değerlendirme beklenildiği için, cümle çeşitlerine ve kelime uzunluklarına göre sınıflandırmak için ayrılan cümleler, kaynak ve insan referans tercüme cümlesi içeren 4 temel farklı iki dilli cümle kaynağından alınmıştır. Cümleler **Bulma (Precision), Duyarlılık (Recall), Bleu, Meteor**, vb. popüler değerlendirme ölçüleri, özellikle Türkçe gibi bitişken dillere uygun olan **Bleu+** kullanılarak meşhur **Google, Bing ve Yandex** gibi çevrimiçi makine tercüme servislerinin çıktı kalitesini ölçmek için karşılaştırılmıştır. Devamında bu otomatik değerlendirme ölçütleri uzman görüşü yorumları ile karşılaştırılarak daha iyi bir otomatik ölçümün nasıl yapılabileceği tespit edilmeye çalışılmıştır.

Anahtar Kelimeler: Çevrimiçi Makine Çeviri Servisi Çıktı Kalitesi, Makine Çeviricisinin Değerlendirilmesi, Tutturma, Tespit etme/ Bulma, Bleu, Meteor, Bleu+

DEDICATION

Dedicated to my dear parents, wife, sister and my little sweet nephew for their endless support and patience during the forming phase of this thesis.

ACKNOWLEDGEMENT

Foremost, I would like to thank God and I would like to express my gratitude to my advisor Dr. Hasan KITAPCI for his encouragement, motivation, guidance, and help on technical issues.

I would like to thank to dear Dr. Ahmet UYAR, Dr. Mete CELIK, Dr. Kadir A. KEPER, Dr. Dogan BULUT**,** Dr. Ali Esat OZMETIN, Dr. Hasan PALTA, Ahmet Selami BALTACI, and Murat Mesut KURAL for their support to my thesis study.

I thank to dear English teachers Ismail GOKBUDAK and Salih YETKIN to their contributions.

I also appreciate to Ahmet Cuneyt TANTUG, Kemal OFLAZER, Zeynep ORHAN and Ihsan Omur BUCAK, for their help and contributions.

I also want to express my gratitude to my colleagues Muhammed Mustafa UNALMIS, Selim DOGAN, Bilge KAGAN DEDETURK, Ayub Rakhman WAKHID, Seyfullah FEDAKAR and Gokhan OZSARI for their supports.

I really thank to the instructors of Foreign School of Meliksah University during human judgment tests.

TABLE OF CONTENT

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF ABBREVIATION

ABBREVIATION

P, Prec: Precision

R: Recall

B: Bleu

BP: Brevity Penalty

B+: Bleu+

M: Meteor

T, TER: Translation Edit Rate

W, WER: Word Error Rate

Avr: Average

Std: Standard

Dev: Deviation

Conf: Confidence

Intv: Interval

MT: Machine Translation

SMT: Statistical Machine Translation

RBMT: Rule Based Machine Translation

HJ: Human Judgment

Smpl: Simple

Cmplx: Complex

Cmpnd: Compound

CC: Complex-Compound

Cmplx-Cmpnd: Complex-Compound

CHAPTER 1

INTRODUCTION

1.1 Motivation of the Research

In the 21st century, called as Information Age, globalization and ever-increasing technological developments affect every aspect of our lives. Citizens of this age need to be able to locate, assess, and represent new information quickly. The Internet become widespread and has created a tremendous content which brings together a big potential for translation. Production of content in multiple languages has become one of the most significant aspects of communicating information.

In recent years, there is a growing challenge trend toward Machine Translation (MT) and its evaluation because of increasing text processing requirements. Machine translation is an automated process by which computer is used to translate text from one natural language to another. The need for machine translation is to overcome language barriers by providing affordable cost and acceptable quality in translating information interested by people.

Machine Translation can be used for different purposes: Meaning, to just extract the essential content of a text; Localization, to translate documentation and help files of enterprises; Communication, to convey the basic content of electronic texts on the Internet, such as web sites, electronic mail and even electronic chat lists; Professional: to contribute to the productivity and efficiency of the work of professional translators.

MT evaluation is a difficult problem. Human evaluation is expensive. However, automatic evaluation is cheap, but not always fair. Human evaluation of machine translation output is an expensive process and inefficient when evaluations must be performed quickly or frequently.

On the other hand, automated evaluations focused on evaluating the correctness of the output, but not the content translation. Nowadays, there are many translation services on the web that aim to translate texts, but nobody knows whether the translation results are completely true or not, except the experts.

There are many written and spoken languages in the world. Most of the internet content is in English and many internet users, who don't know English, want to read and understand the information on the internet. Especially, news readers, students and internet users with less knowledge in English use online dictionaries and translation services to overcome this weakness. However, they don't have any details about translation services' output correctness level. Sometimes many sentence translation meanings are far from real meaning of original text. So using wrong translation outputs of text may cause misunderstandings, even big problems.

Languages, which are source and target language, may belong to different language family in to translation process. In this case translation process steps are increasing and making the translation process complex. At the end, translation process output does not satisfy the real, expected meanings because of lack of some translation steps. If there is a statistical prediction of translation output quality according to sentence types or categories, users are able to decide whether they will use the output of translation services directly or not.

Online translation services can instantly translate between any pair of over fifty languages (such as from French to English). How do they do that? Why does it make the errors that it does? And how can we build something better? Modern automated translation systems like Google Translate and Bing Translator learn how to translate by reading millions of words of already translated text. This research covers a diverse set of fundamental topics from linguistics, machine learning, algorithms, data structures, and formal language theory, along with their application to a real and difficult problem [2].

1.2 Significance of the Research

Nowadays, computational linguistic has very highly interest. And people want to know more information about news in foreign languages, and many people make some agreements with different people who are in different counties that are far away. In Yates's paper, the difficulties of MT were reviewed from the perspectives of the complexity of human language

and translation. Briefly, language is full of exceptions and ambiguities at all linguistic levels. While humans can recognize these extraordinary linguistic features and handle them properly, machines are incapable of performing the same job without adequate human intelligence [1].

Some people want to read news from internet or books. But generally the language of the content is not the same language as the reader's native language. If readers desire to reach meaning of content, they will select one of three ways. First, people may learn other languages which they read about. It may take years. Another one is that they can get help from an expert. It takes weeks and it costs much. Lastly, they can use some dictionaries or automated translation service tools. Last way is the fastest one but may not be reliable. Several tools, free as well, are now available which support translation of text into one or more languages.

In business life, some traders communicate with business chat programs. Many of them are from foreign country. So even so small misunderstanding could result with big business problem [1]. During business and communication, any wrong understanding causes big problems such as loss of money and prestige. So, having information about relief and truth of translation is crucial for those services' users. It is becoming vital what using online translation services directly by users who have insufficient number of people.

Currently there are a large variety of online MT systems to provide nearly instant translation in almost all domains, far faster than human translators can, and free of charge. They serve all popular languages. Their translation quality varies from system to system, from language to language, and from text to text. In the last decade, there were many researches that have grown rapidly on Evaluation of Machine Translation (EMT) under **Natural Language Processing** and **Computational Linguistic** scientific field to measure output quality of MT tools.

1.3 Objective of the Research

This research work will help in providing information to the users about how machine translation evaluation is done and the quality of the online automated translation services. This research work is going to answer the following main questions and related sub questions below:

- Which online translation service provides better quality translation output?
    - What language pair to use to check the quality scores?
    - Are there significant quality scores between sentence structures?
- Which metrics can be used to measure the quality of the automated translation services best correlated with human approach?
    - How to validate automated measurements of services?
    - How to correlate automated measurements of services?

Since bilingual sentences are needed to evaluate the translation services. First aim of the research is to determine some important and useful information about automatic machine translation services. As in similar researches, additional information about the statistical details from the content corpus need to be extracted [3]. However, number of words in a sentence and number of verbs are indeed significant for analyze of a sentence. So it is needed to classify over sentence in terms of sentence types. Because almost every meaning surrounding of verbs.

We report on a horizontal comparison of different online MT systems carried out with a large volume of legal texts. Unlike the approach of evaluating MT quality by human judgment, which is in sharp contrast to the ordinary practice nowadays in the field of MT, our approach adopts the state-of-the-art automatic quantitative MT evaluation technology that has been commonly accepted and widely applied by MT developers and researchers in recent years [1].

1.4 Contributions of the Research

As it is emphasized that “If the evaluation problem will be solved, the translation problem also will be solved” [4], this research focuses on how to evaluate online translation outputs. This research result leads whether online translation services’ are good to be used directly or not by the users; and to give ideas what level online translation services can translate correctly.

This thesis also contains comparison of evaluation methods, both automatic and human evaluation approach, to decide which is useful during translation evaluation process. In addition, this research leads to know whether sentence types or domains have any effects on translation quality by comparison of online web translation services’ outputs.

1.5 Organization of the Research

The following sections provide details about the research. Chapter 2 presents some basic information about my thesis study and I present what I search about it from the literature. Then, Chapter 3 contains the methodology about how the individual steps of research to do in translation evaluation, including collecting data, choosing evaluation metrics, and procedures to follow. Chapter 4 demonstrates the metric results extracted from automated evaluation tools for the online translation services. Chapter 5 summarizes the findings and comparisons between the online translation services. Lastly, in Chapter 6 conclusion and future works are described.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Background

Understanding and getting more benefits is crucial for this study. There is a strong, long and deep infrastructure of this research. The main subtitles of background are on the following with their definition:

- Language (Nature Language)
- Natural Language Processing
- Translation
- Evaluation of Translation

2.2 Language

Language is an effective medium of communication that is used communicates between those who are in the same language acknowledgement field [5]. Spoken languages called natural language [6]. Texts contain sentences made of words. They contain symbols called character. They reflect though to speech. Natural language is unique in being a symbolic communication system that is learned instead of biologically inherited [7].

There are many languages exist in the world. Currently about 6000 languages are on our planet, some spoken by millions, some by only a few dozen people [8]. Languages are named natural and artificial according to their creating duration.

Table 2.2.1: Model of Grammar

Steps	Terms	Content
4	Semantics	Meaning
3	Syntax	Sentence, phrase
2	Morphology	Words, Affixes, Suffixes
1	Phonetics, Phonology	Sounds, Sound Symbols

Model of grammar is consisting of 4 subparts [9, 10]:

- **Phonetic** is concerned with speech, is produced by human mount and at human ear.
- **Morphology** bases on root/stem of word and affix/suffix basic morphemes.
- **Syntactic** is a patterns of sentence in language as well as a phrase or clause consist of words**. Lexical category** interests words type such as noun, adverb, adjective, etc.
- And s**emantic** is a field that studies the meaning of words and sentence [11].

Table 2.2.2: Morphological Analysis

nation
nation-al
inter-nation-al
inter-nation-al-ise
inter-nation-al-is-ation

2.2.1 Spoken and Written Languages in the World

Languages, are generally called natural language in literature, and have many similarity and diversity in the world. So linguists made a classification over them in terms of origins of them.

In the world, internet is a currently good source of information and news. And the language of internet source is English generally. In many countries, people need to know internet information from the real source.

According to the British Council “Language for The Future” report in 2014, languages the most important language for UK’ future was presented on the following:

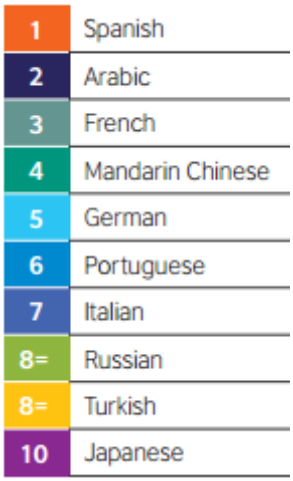


Figure 2.2.2: The Most Important Languages UK's Future [12]

So understanding and managing of these languages are crucial for UK nowadays. In addition, Turkish language is part of the list and in the same level with Russian on the table.

Actually, famous spoken languages are always symbolized with alphabetic characters. So they are both spoken and written languages.

2.2.2 Languages on Internet

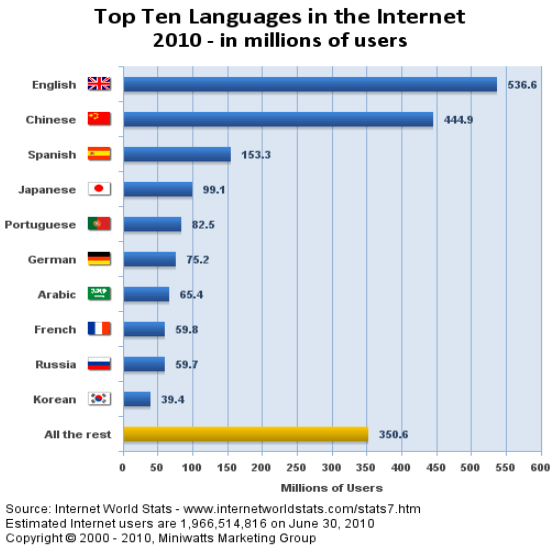


Figure 2.2.2: Languages on Web Preferred by Users

This statistic show that internet users are usually prefers English language on web.

Content languages for websites

Estimates of the percentages of Web sites using various content languages as of 26 April 2013:

Rank	Language	Percentage
1	English	54.9%
2	Russian	6.1%
3	German	5.3%
4	Spanish	4.8%
5	Chinese	4.4%
6	French	4.3%
7	Japanese	4.2%
8	Arabic	3.0%
9	Portuguese	2.3%
10	Polish	1.8%
11	Italian	1.5%
12	Turkish	1.4%
13	Dutch	1.1%
14	Persian	1.0%
15	Czech	0.6%
16	Swedish	0.6%
17	Indonesian	0.5%
18	Greek	0.4%

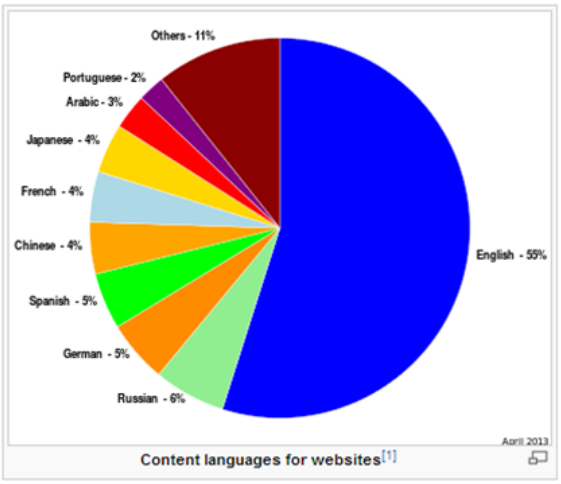


Figure 2.2.3: Content of Languages on Websites

However Content of web sites is mostly written in English language in the world, Turkish language is also on the web significantly.

2.3 Basic Terminology for Language

Languages consist of texts. Texts are occurred in paragraphs. Sentences contain words and words are formed by characters/symbols.

There are generally two approaches to define structure of sentences:

1. **In terms of sentence type**
2. **In terms of clause type**

Table 2.3.1: Sentence Types in terms of Punctuation [13]

Sentence Type	Explanation
Statement	Declarative
Question	Interrogative
Exclamation	Exclamatory
Command	Imperative

The first approach, taking into account of types, is generally related with meaning and punctuation. In the second approach, sentence classification is based on structure of sentence. Structure is shaped by number of clause and their relations.

Table 2.3.2: Sentence structure

Structure Title	***Clause Properties***	***Sample***
Simple	*Only 1 independent clause*	*The dog barked.*
Compound	*At least 2 independent clause*	*The dog barked* ***and*** *the cat yowled.*
Complex	*At least 1 independent and 1 dependent clause*	*The dog* ***that was in the street*** *howled loudly.*
Compound and Complex	*At least 2 independent and 1 dependent clause*	***As the dog howled***, ***one cat sat on the fence, and the other licked its paws***.

This approach gives us a list about sentence class separated with clause statement. There are two clause types [14]:

1. **Independent Clause**
2. **Dependent clause**

Complex sentence is consists of a combination of an independent clause and a dependent clause. An example with a ***relative clause*** as the dependent clause:

Example 1: The dog **that was in the street** howled loudly.
Example 2: A student **who is hungry** would never pass up a hamburger.

An example with a *subordinating conjunction* creating the dependent clause (note the various positions of the dependent clause):

Ex: End: The dog howled **although he was well fed**.
Ex: Front: **Because the dog howled so loudly**, the student couldn't eat his hamburger.
Ex: Middle: **The dog, although he was well fed**, howled loudly**.**

Compound sentence is consists of two or more simple sentences (Independent Clause) joined by:

- **a comma followed by a coordinating conjunction (*and, but, or, nor, for, yet, so*)**: Ex: The dog barked**, and** the cat yowled**.**
- **a semicolon**: Ex: The dog barked**;** the cat yowled.

Compound-complex sentence is consists of a combination of a compound sentence and a complex sentence [15].

Ex: **As the dog howled**, ***one cat sat on the fence*, and *the other licked its paws***.

During separating sentence structure there are some confusion and harnesses.

1- Compound subject and verb statement using with "and" conjunction in simple sentence versus compound sentence
2- Complex sentence without subordinates such as "that, who, which"
3- Adverb and noun-verb confusion

So sentence cannot be automatically separating with automatic methods and they must separate by hand manually. Therefore/hence there may be some classification errors.

2.4 Natural Language Processing

Natural language processing (NLP), is a highly interest scientific field nowadays. Because, management of increasing numbers of languages is very crucial.

NLP is a branch of artificial intelligence that deals with analyzing, understanding and generating the languages that humans use naturally in order to interface with computers in both written and spoken contexts using natural human languages instead of computer languages [16].

NLP is a branch of information machine science that deals with natural language information [17]. There are many NLP fields [18] such as:

- Text Classification
- Information Extraction
- Information Retrieval
- Machine Learning

- Question Answering
- Word Semantic
- Machine Translation
- Evaluation of Machine Translation

Especially my interest is evaluation of machine translation scientifically natural processing field. Because output correctness rates of machine translation systems are uncountable flexible, insufficient, unstable, and unexpected level. One of the main NLP areas is translation between languages and their evaluation. Both of them can be made by human and machine now.

2.5 Translation

Translation is a process of translating text from one language into another [19]. In order to translate, additionally reordering positions of words is needed. Also determining chunks and phrases is mostly significant.

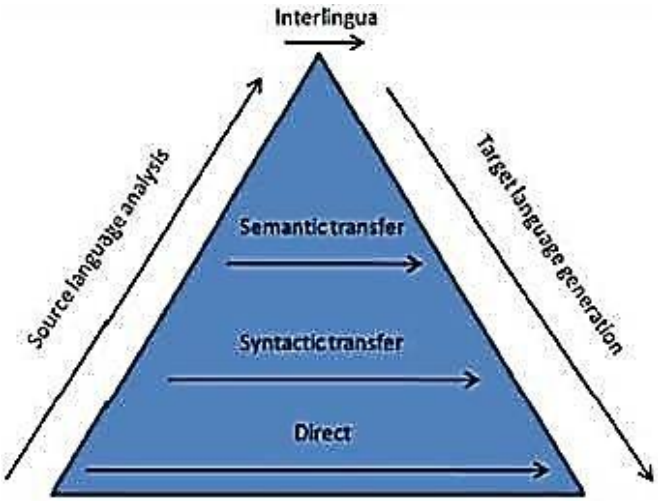


Figure 2.5.1: Translation Steps

Generally, translation on language is made by transformation of meaning on different level. There are mainly two approaches of translation such as human translation and machine translation.

2.5.1 Human Translation

Translation between languages is one of the most significant requirements of linguistic. Since its high accuracy, human translation is accepted and preferred more reliable than other translation types [20].

Human translation is the perfect one of translation approach, but it has many various solutions in terms of expert's mind, approach and culture. However all of these useful side of human translation, it is expensive, it takes time and, more human labor is required than fast one, machine translation with shortcomings.

2.5.2 Machine Translation

Machine translation is a one of the most challenges / research areas of computational linguistics in computer science. It was adopted to communicate the texts from one language to another [5]. It also called "automatic", "computer-aided" translation that is made by computers/machines. There are many automatic language translation approaches. Different methods of machine translation are presented in the followings:

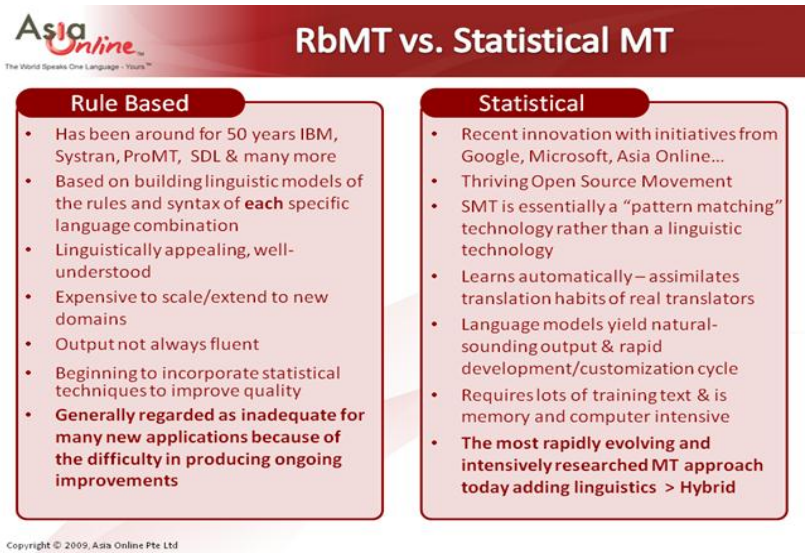


Figure 2.5.2: Comparison of SMT and RBMT

The figure below gives specifications of rule-based and statistical machine translation approach. And more approaches used in literature are on the following table:

Table 2.5.1: Machine Translation Types

MT Types	Definition
Rule-Based	Rule-Based Machine Translation systems use large collections of rules, manually developed over time by human experts mapping structures from the source language to the target language [21].
Statistical-Based	Statistical Machine Translation systems use computer algorithms to produce a translation that looks best statistically from millions of permutations [22].
Hybrid-Based	In order to address quality and time-to-market limitations, many Rule-Based Machine Translation developers are augmenting their core technology with Statistical Machine Translation technology to create 'Hybrid Machine Translation' solutions [22, 23].
Example-Based	It also variant of corpus based machine translation type that databases of already translated examples are used for matching against new input and proper samples are extracted after recombination with analogical manner to determine the correct translation [24].

Also there are many other MT types such as phrased-based MT, knowledge-based MT, etc., but the most common MTs in use are on the table 2.3.1.2 above [25].

Online Machine Translation Services and Their Limitations

There are many online translation services on web. But their abilities are different. Some of them have character/symbol restrictions. They almost have secondary paid-based translation. Some of them are allowing only word translation and some of them have both word and sentence translation facility. Almost all of the most popular services on web with their properties on the followings [26]:

Table 2.5.2: Comparison of MT Application

Name	Platform	Price
Asia Online	Windows, Linux, Web	Trial Demo
Apertium	Unix	No Fee Required
Anusaaraka	Unix	No Fee Required
IBM	Cross-platform	Commercial
OpenLogos	Windows, Linux	No Fee Required
Moses	Cross-platform	No Fee Required
NiuTrans	Cross-platform	No fee required
Google Translate	Web application	No fee required
Bing Translator	Web application	No fee required
SAIC	Windows, Linux, Web, i0S	Depends on configuration
SYSTRAN	Web application	Commercial
GramTrans	Web application	No Required Fee
Promt	Web application	Commercial
SDL	Web application	Commercial, Trial Demo
Babylon	Web application	Commercial, Trial Demo
WorldLingo	Web application	Trial Demo
IdiomaX	Mobile	Depends on configuration
Transsoftware	Windows	Commercial
Yandex	Web application	No Fee Required

They are all sentence supported services. Some other services web based translation tools can only translate words or word groups and they give same outputs with the most popular online translation services like Google, Bing, etc. Some of them have APIs (Application Program Interface) to integrate programs to use these services easily.

2.6 Evaluation of Translation

Evaluation of translation is a determine process what your credentials are worth in terms of the educational system [27]. There are two mainly subsection of evaluation; human (manually) and machine (automatically) evaluation. Somewhere human evaluation called human judgment. Evaluation is based on matching similarity of two texts in same language. Similarity is based on bath word, phrase and meaning.

2.6.1 Human Evaluation (Human Judgment)

Expert can measure correctness rate of translation results according to their acknowledgement. This is also named "Subjective" Evaluation [28]. According to paper, there is a 3 way of classification of state detection of sentence. Firstly, semantically and syntactically correct with respect to reference/s sentence. Then semantically correct, syntactically incorrect. And lastly, in terms of both approaches incorrect. It means that if a sentence syntactically incorrect, it cannot be correct semantically.

Human evaluation is also called observation or empirical based evaluation. There are two main human evaluation approaches such as ***adequacy*** and ***fluency*** [29]. There are two common approaches of human evaluation in literature on the followings:

Table 2.6.1: The Rating Levels for Evaluation [30]

1	**Unacceptable**	Absolutely not comprehensible and/or little or no information transferred accurately.
2	**Possibly Acceptable:**	Possibly comprehensible (given enough context and/or time to work it out); some information transferred accurately
3	**Acceptable:**	Not perfect (stylistically or grammatically odd), but definitely comprehensible, AND with accurate transfer of all important information
4	**Ideal:**	Not necessarily a perfect translation, but grammatically correct, and with all information accurately transferred

This first one of evaluation rating approach's levels can be extended in range of 5 individually and especially steps with their explanations in parts of ***fluency*** and ***adequacy*** [31].

Table 2.6.2: Human Evaluation Criteria and Steps [31]

	Fluency	Adequacy
1	Incomprehensible	None
2	Disfluent language	Little meaning
3	Non-native language	Much meaning
4	Good language	Must meaning
5	Flawless language	All meaning

Second approach is on the table above where ***fluency*** measures whether a translation is fluent, regardless of the correct meaning, and ***adequacy*** measures whether the translation conveys the correct meaning, even if the translation is not fully fluent [32].

2.6.2 Automatic Machine Evaluation

It is called objective measurement [33] that is based on machine aspect accuracy. There are many methods to evaluate sentence according to outputs of machine translators.

2.6.2.1 Automatic Machine Evaluation Methods

Below, the set of lexical measures are described used in this work, called **Lexical Similarity,** grouped according to the type of measure computed. There may some score differences between original formula and calculated measurement by hand and by automatic tools. And there are some samples given on the below to explain metrics much more.

Table 2.6.2.1.1: Automatic Machine Evaluation Methods Class

Edit Distance	Precision	Recall	F-Measure	NGRAM
WER	PRECISION	ROUGE	GTM	Bleu
PER	NIST	RECALL	METEOR	Bleu+
TER	BLEU		F1-Measure	NIST
	Bleu+		OVERLAP	

Edit Distance Based Metric

WER (Word Error Rate) is used as a precision measure [34]. This measure is based on the Levenshtein distance (Levenshtein, 1966) the minimum number of substitutions, deletions and insertions that have to be performed to convert the automatic translation into a valid translation (i.e., a human reference) [53]. The WER formula on the following:

$$WER = \frac{S + D + I}{N}$$

, where S is meaning of substitution, D is corresponding to Delete, I character is representing to Insertion, N means number of words in reference texts and Word Accuracy Score equals (1-WER). There is a single sample to explain how WER score calculated [54].

Ex:

R:	SAUDI ARABIA	***	***	denied	THIS WEEK	information	published	in	the	AMERICAN	new	york	times
C:	THIS WEEK	THE	SAUDIS	denied	***	***	information	published	in	the	***	new	york times
Ev:	S	S		I	I		D	D			D		

There are 2 Substitution, 2 Insertion and 3 Deletion. So:

Word Accuracy = 1- 7/13 = 6/13 = 46.1%

PER (Position-independent Word Error Rate) has a similar type of WER metric but a shortcoming of the WER measure is that it does not allow reordering of words. In order to overcome this problem, PER compares the words in the two sentences without taking the word order into account. Word order is not taken into account [53].

TER (Translation Edit Rate) measures the amount of post-editing that a human would have to perform to change a system output so it exactly matches a reference translation. Possible edits include insertions, deletions, and substitutions of single words as well as shifts of word sequences. All edits have equal cost. There is only one extra operation, shift, makes TER different from WER [35]. The TER formula is same with WER but S character is corresponding to "Shift" operation [35].

Ex:

R:	SAUDI ARABIA	denied	THIS WEEK	information	published	in	the	AMERICAN	new	york	times.		
C:	THIS WEEK	THE	SAUDIS	denied	information	published	in	the	***	new	york times.		
Ev:	SHFT		S	SHFT					D				

Two shifts, one substitution and one deletion. So Word Accuracy rate equals (1- 4/13) 9/13 = 69.2%

Lexical Precision

Pl stands for Lexical Precision, it computes the min-intersection of items (tokens) in the reference and the candidate divided by the items in the candidate. **Precision** is basically morphological compare based method. It matches number of same words in sentences opposing to number of candidate (machine translation) sample [31].

Ex:

Ref:	SAUDI ARABIA **denied THIS WEEK information published in the** AMERICAN **new york times.**	13
Cand:	**THIS WEEK** THE SAUDIS **denied information published in the new york times.**	12

Exactly/Surface Matching is 10. So Score of Precision is 10/12. It equals 83.3%.

BLEU (Bilingual Evaluation Understudy) is used to accumulated and individual BLEU scores for several n-gram lengths (n = 1::: 4, default is 4). Default is accumulated BLEU score up to 4-grams and smoothed as described by Lin and Och [37, 38].

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

Figure 2.6.2.1.1: Bleu Formula

A more recent idea is that matching words sequence in right order has high scores than out of order [39]. In terms of this idea, a simplification which named "***BLEU***" has been described. In that description, there is a measurement of syntactic similarity between a candidate a reference by counting the number of matching n-grams for $1 \leq n \leq 4$ [37]. Here is an example to show how bleu calculate its score:

Ex:

SAUDI ARABIA **denied THIS WEEK information published in the** AMERICAN **new york times.**
THIS WEEK THE SAUDIS **deinied information published in the new york times.**

Table 2.6.2.1.2: N Gram Sequence Sample for Bleu Measurement

1-gram P.: 10/12	Matching Words one by one
2-gram P.: 6/11	This week, information published, published in, in the, new york, york times
3-gram P.: 3/10	information published in, published in the, new york times
4-gram P.: 1/9	information published in the

Average logarithm of n-gram Precision: 35% and BP: e^(-1/12) equals 92%. **Score of BLEU** equals BP x Average N-gram Precision Logs. So it is **32.2%.**

Bleu+ method [40] method is a fine grained version of Bleu. This method is specifically for agglutinative languages such as Turkish, Hungarian, Finnish, etc. It calculates word root/stem and suffixes similarity level by suffix based Levenshtein distance method. Bleu+ plus Formula is in the following:

Table 2.6.2.1.3: Bleu+ Approach Formula Basics

$$S(w_i, w_j) = S_{\text{root}}(w_i, w_j) \times S_{\text{morph}}(w_i, w_j)$$

BLEU+ tool provides a graphical user interface through which various options can be set. Using with it has a synonym list, similar word or phrase meanings are accept almost same. Here is an example which is same with previous bleu sample on the following:

Ex: Stem/Suffix Matching: 11 **(SAUDI -> SAUDIS)**

- 1-gram prec.: 11/12
- 2-gram prec.: 6/11
- 3-gram prec.: 3/10
- 4-gram prec.: 1/9

SBLEU = BP x Avr. Log. of n-gram Precision X = **33.06%**

NIST (National Institute of Standards and Technology) is used accumulated and individual NIST scores for several n-gram lengths (n = 1::: 5, default is 5). Default is NIST score up to 5-grams [41].

Lexical Recall

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) has eight variants that are available. It is basically based on recall and different n gram types of combination [38].

RI stands for Lexical Recall, it computes the max-intersection of items (tokens) in the reference and the candidate divided by the items in the reference [42].

Ex:

Ref:	SAUDI ARABIA **denied THIS WEEK information published in the** AMERICAN **new york times.**	13
Cand:	**THIS WEEK** THE SAUDIS **denied information published in the new york times.**	12

Exactly/Surface Matching is 10. And Score of Recall equals10/13. So it equals **76.9%.**

F-MEASURE based methods, almost generally used as F_1-MEASURE, is also called f-score or f-measure that is a measure of accuracy unit [43]. It is based on precision and recall which are balancing together in terms of β (β ∈ positive real numbers). The general formula of f-score on the following:

$$F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

Traditional equal balanced F-score formula is above when β equals 1.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

This means that F1 score is harmonic average of precision and recall. It is used for to obtain balanced ratio.

METEOR is another sentence similarity evaluation method, also the closest one to human judgment, which can measure words by calculating harmonic mean of precision and recall [44]. Additionally, it tokenizes sentence to prepare for evaluation by removing dashes between hyphenated words and removing full stops in acronyms/initials. It also compares sentence and words on stem, synonym and paraphrase level [45].

$$Fmean = \frac{PR}{\alpha \cdot P + (1-\alpha) \cdot R}$$

$$Penalty = \gamma \cdot \left(\frac{\#chunks - 1}{\#unigrams\_matched - 1} \right)^{\beta}$$

$$Meteor = Fmean * (1 - penalty)$$

Original weights: $\alpha = 0.9,\ \beta = 3,\ \gamma = 0.5$

Figure 2.6.2.1.2: Meteor Formula

Ex:

SAUDI ARABIA denied THIS WEEK information published in the AMERICAN new york times.
THIS WEEK THE SAUDIS denied information published in the new york times.

P = 83.3% **S**R = 76.92%

Fmean = 0.77 (Recall Weighted)

Penalty: 4 chunks, 10 matched words = 0.5 x (4-1)/(10-1) = 0.52

Meteor = Fmean x (1-Penalty) = **36.40%**

2.6.2.2 Automatic Metric Evaluation Tools

There are little, free but effective evaluation of bilingual data corpus tools on the following:

Table 2.6.2.2.1: Evaluation Tools

Title	Type	Supported Methods
Asia Online	Executable	Bleu, F-Measure, TER, Meteor
Asiya	Web Application	Precision, Recall, TER, WER, Per, F-Measure, Meteor, Bleu, etc.
Bleu+	Executable	Bleu+

Figure 2.6.3: Asia Online Language Studio Automatic Metric Evaluation Panel

Asia online language tool is a downloadable and executable program that allows evaluating text similarity in trial version. It contains Bleu, F-Measure, TER, Meteor metrics. It can evaluate sentence list together only.

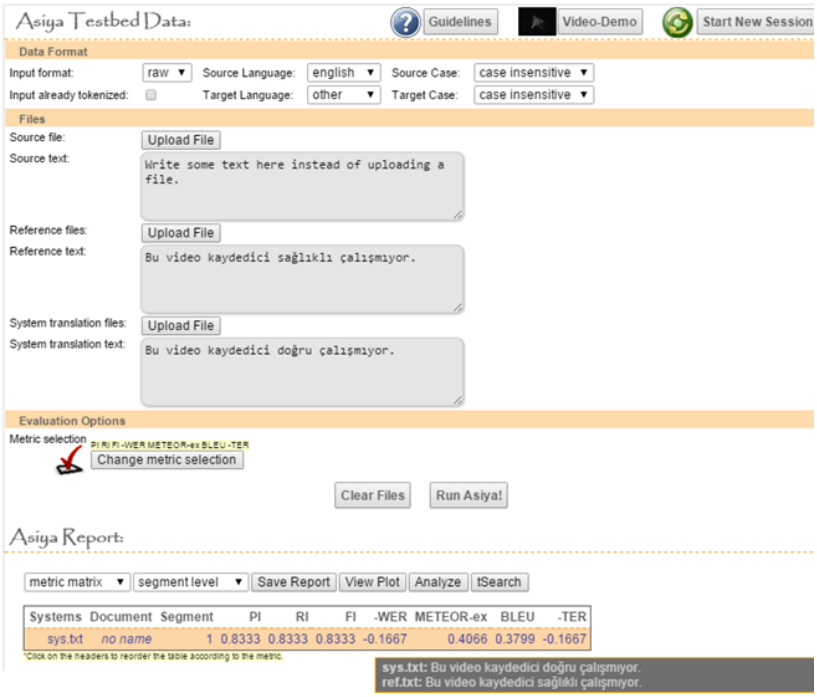


Figure 2.6.2.2.1: Asiya Evaluation tool

Asiya evaluation tool has an online interface. It contains a lot of evaluation metric such as Precision, Recall, F-measure, TER, WER, and Bleu, etc. It is available from everywhere and can evaluate sentence list individually.

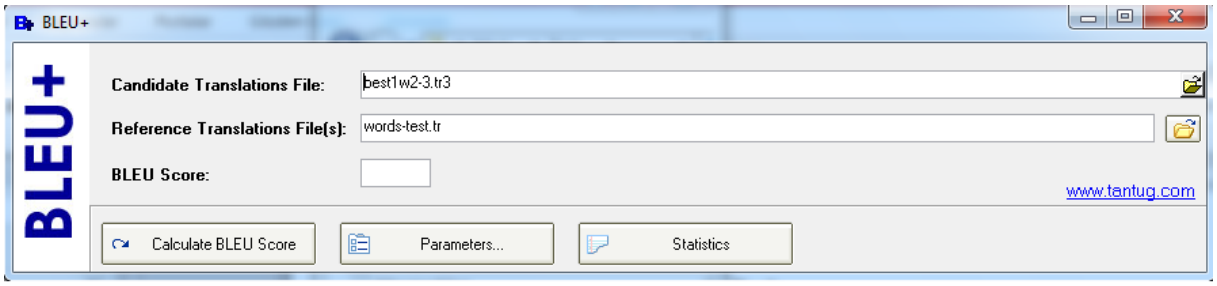


Figure 2.6.2.2.2: Bleu+ MT Evaluation Tool

Bleu+ is an extended version of Bleu created by Tantug [38]. It calculates root and suffixes individually. And if there is more than 70% similarity that word is considered as matching word. This approach is more flexible than bleu.

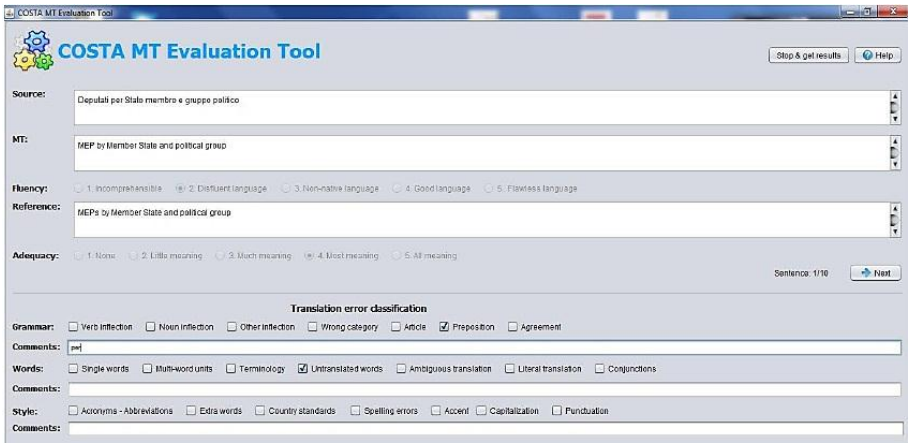


Figure 2.6.2.2.3: Costa Human Evaluation Tool

Costa Mt evaluation tool allows to be ranked sentences individually in range of 1 to 5 in terms of adequacy and fluency criteria by human. And it allows adding some extra comment.

2.7 Related Works

There are many researches on scientific and computational linguistic area. The main surveys are the followings:

- Machine Translation and Its Evaluation
- Evaluation Methods of Sentence Similarity Measurement

Table 2.7.1: Literature review summary [55-59]

Title	Languages	Dataset	Result
A Short Guide to Measuring and Comparing Machine Translation Engines	**English to French**	**3 different and 1 combined bilingual reference corpus**	**Google > Bing > Systran**
			Bleu> Others
Subjective and Objective Evaluation of English to Urdu Machine Translation	**English to Urdu**	**Sample Sentence**	**ATEC> METEOR**
Two Phase Evaluation for Selecting Machine Translation Services	**Japanese to English**	**data-driven classification, 300 -> 6 groups**	**j-Server > Google > Translution**
			Bleu > NIST > WER
Which online translation is best in Spanish to English translation?	**Spanish to English**	**1. corpus**	**Bing > Babylon > Google > Free Translation > Prompt**
		2. corpus	**Google > Bing > Prompt > Free Translation > Babylon**
BLEU+: A Tool for Fine-Grained BLEU Computation	**English to Turkish**	**Sample Sentence**	**Bleu+>Bleu**

Scientist tries to find out the state of many different service output qualities comparatively by using with different corpus and metrics. They believe that this approaches help machine translation engines to improve their self. In terms of language specifications corpus give different results by using with different metric.

Early approaches to compare text to detect similarity rate with calculating number of matching words [46]. Some basic evaluation methods such as ***Precision***, ***Recall*** and ***F-measure*** show significantly higher correlation with human judgments over 728 English – Arabic bilingual sentences Corpus with six different translations which of two are coming from machine translation and four of them are coming from reference (human) translation [47].

There is another survey about evaluation of 5 popular web-based MT systems in empirical usability factors [48]. This study contributes into development of on-line MT services to enhance their design to be useful real users.

There is an extensive survey [49] about comparison among these methods' translation accuracy evaluation rate from one language to another one. Especially, Chinese, Arabic, and English Languages are used to measure evaluation of translation. BLEU and NIST are strong, given high rate evaluation methods of machine translation. They are still the best general choice for training model parameters. Models trained using n-gram based metrics, BLEU and NIST, are more robust to being evaluated using the other metrics. It is determined that Meteor works reasonably well for Chinese but is not good choice for Arabic.

The RYPT based metric which directly makes use of human adequacy judgments of substrings, would obtain better human results than the automated metrics presented here [51].

Since using BLEU and NIST produces models that are more robust to evaluation by other metrics and perform well in human judgments, we conclude they are still the best choice for training [49].

There are many automatic evaluation methods to measure similarity rate of sentences. But researches are shown that methods below are most successfully and meaningful for evaluation of English and Turkish languages:

Table 2.7.2: Suitable Evaluation Methods vs. Languages Table

Methods vs. Languages	En	Tr
WER	**	
TER	**	
METEOR	***	
BLEU	***	*
PRECISION	**	*
RECALL	**	*
BLEU+	-	***

The number of asterisk means that the scientific papers related with this evaluation subject which contain similarity evaluation metric consider about these metric densely.

And there is a sentence evaluation tool for similarity comparison such as Asia Online Language studio, Asiya Online evaluation service and Costa MT evaluation tool [31]. Excluding last one are automatic evaluation tool with more than one automatic evaluation methods like Bleu, TER, Meteor, etc. Costa MT evaluation tool helps us to evaluate sentence by using criteria in terms of human judgment manually.

CHAPTER 3

METHODOLOGY

In this study, a number of the most commonly used online MT services was examined and compared with their translation performance on legal texts, a text genus of particular importance to newspaper readers and internet users. Methodology of this thesis is consisting of 4 steps of Research on the following:

3.1 Selection of Online MT Services

We highly interested in and focused on Turkish and English languages. So the services we can use should support these languages. Services which we selected should also be free accessible and cross platform (web based) at same time to reach easily without any login. The table is given below about online translation services and their attributes [26].

Table 3.1.1: Translation Services over Internet

Name	**Platform**	**Price**	**TR&En Support**
Asia Online	Windows, Linux	Depends on configuration	**Yes**
Apertium	Unix	**No Fee Required**	No
IBM	**Cross-platform**	Commercial	No
NiuTrans	**Cross-platform**	**No fee required**	No
Google	**Cross-platform (Web application)**	**No fee required**	**Yes**
Bing	**Cross-platform (Web application)**	**No fee required**	**Yes**
SYSTRAN	**Cross-platform (Web application)**	Commercial	**Yes**
Promt	**Cross-platform (Web application)**	Commercial	No
SDL	**Cross-platform (Web application)**	Commercial, **Trial Demo**	**Yes**
Babylon	**Cross-Platform**	Commercial, **Trial Demo**	**Yes**
WorldLingo	**Cross-Platform**	Commercial, **Trial Demo**	**Yes**
Transsoftware	Windows	Commercial	No
Yandex	**Cross-Platform**	**No Fee Required**	**Yes**

Services compatible for my requirement are on the following with their properties and limitation [51, 52, 53]:

Table 3.1.2: Services Used in Tests

Service / Property	Limitation	# Lang. Sup.	MT Types	Instant Mode	Pronunciation
Google	Unknown	81	SMT	√	√
Bing	5.000 Symbol	44	RBMT & SMT	√	√
Babylon	300 character	30	RBMT		
World Lingo	500 words	33	RBMT		
SDL	500 Words	43	KBMT	√	
Yandex	10.000 char.	43	SMT	√	√

As seen that over 6 popular TR EN supported online machine translation services have some properties such as languages, character limits and MT types. Interface of those services are follows:

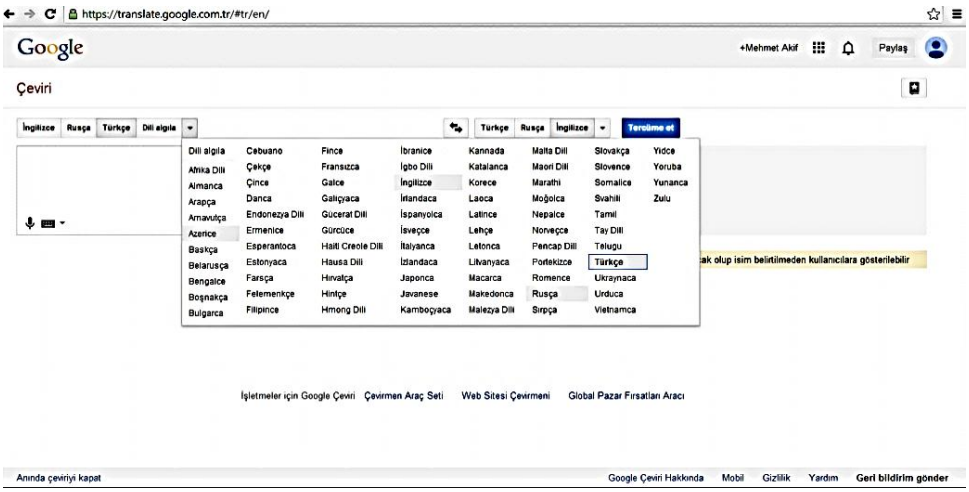


Figure 3.1.1: Google Translation Service Interface

During obtaining words or sentence, Google allows selection/changing synonym of words or words group / expression, but Bing, Yandex and others not yet. Manually matched word tracking can be made by cursor movement on words at Google with yellow signing. Additionally, documents and web sites with URLs can be used as input for translation. Google translation services support 81 languages against and it accepts many large size of text together. It presents some functional use such as input text with microphone and uploads from text file. It allows hearing pronunciation of source texts and translation.

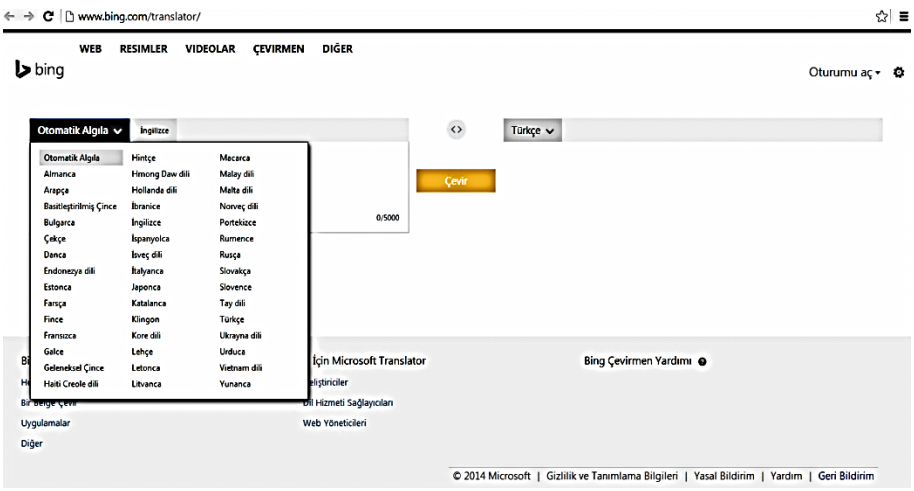


Figure 3.1.2: Bing Translation Service Interface

Bing translation services, also other Bing services, are in part of Microsoft products. It shows written character counts over 5000. Then manually matched words tracking can be made by cursor movement on words at Bing with yellow signing. You can hear text pronunciations also. Bing translation web services has a recently limitation about 5.000 characters called symbol. It also supports 44 languages.

Figure 3.1.3: Babylon Translation Service Interface

Babylon also provides a downloadable program. Trial version is available for online. Short texts are allowed for translation.

Figure 3.1.4: World Lingo Translation Service Interface

World lingo translation service is almost same as a Babylon Service.

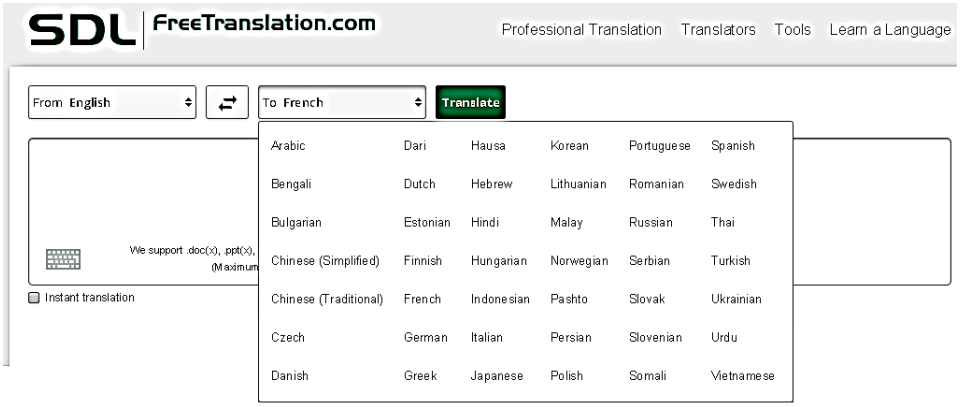


Figure 3.1.5: SDL Translation Service Interface

Another service, SDL, is similar to Babylon. “Instant Translation” selection mode is additionally available.

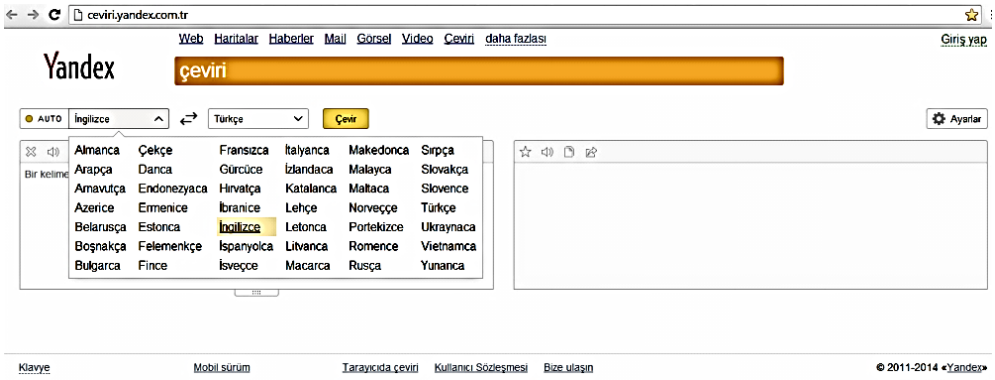


Figure 3.1.6: Yandex Translation Service Interface

Manually matched words tracking can be made by cursor movement on words, at Yandex with yellow signing. You can also hear text pronunciations. Yandex has 10.000 character limitations. And it has about 43 language support.

These services (Google, Bing, and Yandex) have also an API (Application Program Interface) to obtain translation easier and faster but we did not prefer using API due to integration process because of usage reality.

In order to obtain translation outputs fast and easily, we ignored Babylon, World Lingo and SDL web services due to their low character and word limitation. It makes difficult translation of big data to evaluate in my thesis. At the end we decided to use Google, Bing and Yandex translation web services only. These services are more suitable for my study.

3.2 Bilingual Data Corpus Collection

During my research, I have not founded any bilingual corpus separating in terms of domain. So I collected legal sentences from the following sources. Various type of source for bilingual data is need to variety. Because online translation service users put sentence in different domain such as academic, news, daily, historical, etc.

Table 3.2.1: Corpus Sources

Bilingual Corpus Source	Explanation	Simple	Compound	Complex	Compound-Complex
Osym.gov.tr - Exam Center	**ÜDS-KPDS-YDS Translation Question and Answer**	4	18	44	8
Manything.com - tatoeba.org/	**Bilingual sentence pairs archive**	6	-	-	30
Yeminlisozluk.com	**Bilingual sentence pairs archive**	1	-	-	-
lonweb.org	**Bilingual sentence pairs archive**	1	-	-	-
News, Education Documents – Expert Translations	**Human Translators are native Turkish speaker English Teachers.**	6	13	4	2
wikipedia.org/	**Bilingual sentence pairs samples**	2	-	-	-
Historical Book	**Bilingual Book**	25	17	2	-
Assay, Thesis Abstract	**Bilingual Academic Abstract**	5	2	-	-
Totally		**50**	**50**	**50**	**40**

Because there are no enough bilingual corpus which contains compound-complex sentences, we decide to reduce all number of sentence train test to 50 equally except compound-complex sentence structure. There are only 40 sentences in compound-complex part.

It is interesting and expected information that almost every English source text word number is higher that Turkish reference human translation. It is prove that an expression in Turkish can explain with more than one word. In other word, an expression consist of more than one word can express in one word with root and suffixes.

Table 3.2.2: Number of Word Comparison of Train Set on Source-Reference Corpus

Much # of word	Count of Longer sentences in terms of # of word
Source English Text	183
Reference Turkish Human Translation	3
Same number of word	4
Totally	190

This table means that texts are longer in English language than Turkish language in terms of word number. These results prove that some English words represents as a suffix in Turkish language. Since, Turkish texts are generally shorter than English texts.

Table 3.2.3: Number of Words Statistics in Source-Reference Sentence Structures

Sentence Structure	Min	Max	Avr.
Simple	4.0	23.0	11.4
Compound	10.0	33.0	22.5
Complex	9.0	37.0	20.4
Compound-Complex	15.0	36.0	24.1

These numbers are coming from the number of word measurement on original source text to realize length of sentence in terms of structure.

3.3 Classification of Data

Data, coming from bilingual corpus, is separated two parts. First part was used for training test to obtain average values to use assumptions/predictions/estimation. Second part is used for verification test.

Data is called corpus, too. Corpus has bilingual sentence pairs in **Turkish** and **English**. Sentences were classified manually in term of instructions into 4 parts: **Simple**, **Complex**, **Compound** and **Complex-Compound**. We have selected these sentence structure based classifications because there is no study we have ever seen yet over papers up to time. So we decided to focus on this field to innovative study.

Table 3.3.1: Bilingual Sentence Structure Sets Distributions

Sent. Str.	Train Set	Automatic Verification Test	Human Judgment Test
Simple	50	15	16
Complex	50	15	10
Compound	50	15	4
Complex-Compound	40	10	2
Totally	190	45	32

Totally 32 sentences from inside train set to evaluate especially by human judgment since to decrease Expert evaluation process duration time also compared with automatic metrics.

Table 3.3.2: Word Statistic over Corpus

	Train Test Set					Verify Test Set				
Word Statistic		**Source-En**		**Reference-Tr**			**Source-en**		**Reference-Tr**	
Type /Attribute	*# of Sent.*	*unique*	*total*	*unique*	*total*	*# of Sent.*	*unique*	*total*	*unique*	*total*
simple	50	280	570	283	423	15	110	145	102	106
complex	50	596	1129	648	832	15	74	119	64	95
compound	50	479	1018	522	769	15	57	102	54	82
complex-compound	40	481	963	519	707	10	184	300	171	232
total	190	1836	3680	1972	2731	55	425	666	391	515
Unique Word Rate		**49.8%**		**72.2%**			**63. 8%**		**75.9%**	

Because of variety of suffixes, in Turkish language, exact unique words counts, determined surface matching, are bigger than English language nationally.

3.4 Evaluation Methodology

To judgment of sentence via human experts easily, forceless, and quickly, we got sentence thorough reducing sentences from 100% to about 13%. So we will evaluate a subset of 245 sentences which is about 32 to verify with human approach easily. There are 2 mainly evaluation aspect of similarity, such as, Machine and human evaluation.

Inside overall 245 sentences selected 50 sentence pairs are used for train test and randomly selected 15 sentences are used for verify test to comparison. There is an exception only to complex-compound sentences: 40 sentence for train test and 10 sentence pairs to verify test because of limited source for complex-compound sentences in corpus which we collected.

Table 3.4.1: Number of Sentence Distribution for Training and Verify Tests

	Total	Train Test	Verify Test
Simple	65	50	15
Complex	65	50	15
Compound	65	50	15
Complex-Compound	50	40	10
All Sentence	245	190	55

Results are compared respectively correlation among training and testing scores.

3.4.1 Machine Evaluation (ME) Step

Sentence similarity is checked by using with algorithms; especially they called methods, to measure similarity of sentences objectively in terms of many aspects. We selected a subset of evaluation which is **TER**, **Precision, Recall, Bleu, Meteor** and **Bleu+** evaluation methods because they are popular, easy to access. Their results are between in range of 0 and 1 or 0 and 100. I normalized all rates in range of 0 and 100.

Corpus was separated into 2 parts: Train and automatic validation set .First set is used for to obtain automatically average and range scores from bilingual corpus about quality of service translation by **190 sentences** total. Almost for every sentence structures such as simple, compound, complex and compound complex, 50 sentences are taken excepted compound-complex structure. That subpart contains only 40 sentences since there are no more sentences about this structure taken from relevant bilingual corpus. And second part is used for automatically validation by **55 sentences** consist of 15 simple, compound, complex and 10 compound-complex sentences

3.4.2 Human Evaluation (HE) Step

Although human evaluation also called judgment is not same with auto metrics one to one, there are some basic criteria to evaluate sentence similarity. As mention at section 2.6.1 we preferred **Adequacy** and **Fluency** criteria which declared in the previous chapter 2. Their evaluation rates are in range of from 1 to 5. But “All meaning” and “Flawless language” (5) options and “None” and “Incomprehensible” (1) options can be prefer in human evaluator mind while that candidate sentence is not the best or worst translation. But Costa Human Evaluation tool with 20% in first step. So there is no similarity you can select “1”. None of

meaning, tool gives you 20% similarity rate. Costa MT tool designers might assume that the worst sentence can be reflect a few opinion about source text.

There are **32 sentences** in the human judgment corpus. Sentence are well balanced by sentence structure in terms of big 830 corpus. So 8 simple, 5 complex, 2 compound and 1 compound-complex sentence are collected inside to one human judgment evaluation set. Each of From Turkish to English and English to Turkish datasets has same number of sentence. And 3 of translation services' outputs were evaluated by 8 different native Turkish speaker English teachers.

3.5 Statistical Computation and Representation

I demonstrate meaning of results with Z distribution confidence interval estimation is on the following formulas:

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \le \mu_X \le \bar{X} + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right) = 1-\alpha$$

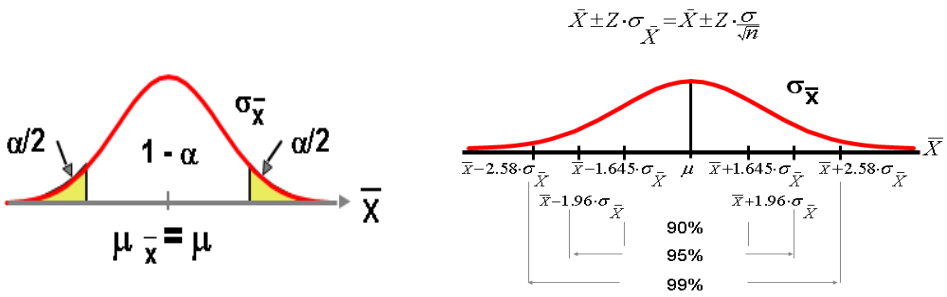


Figure 3.5.1: Confidence Interval Calculation Basics

These basics are coming from statistic science. Scientists when they desire to compute density range with ratio, they calculate confidence interval to overreach of standard deviation. So confidence interval calculation helps to

$$\bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \quad \text{upper bound}$$

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \quad \text{lower bound}$$

Figure 3.5.2: Confidence Interval Formulation

, where X represents mean value of samples, σ (sigma) represents standard deviation and n represents number of sample. 1.96 is confidence interval coefficient for 95%.

3.6 Evaluation Tools

The tools we have selected only Asiya web based tool [54] and Bleu+ tool [40] to get automatic assessment of quality because Asiya evaluation tool is contain all method of Asia online executable evaluation tool and anymore. Although Asia online is offline executable program. So it needs installation. Asiya is an online reachable tool and it allows evaluation of sentence both individually and collectively. So we selected Asiya evaluation tool.

To judgment of human mind, there is a useful tool; its name is Costa MT [31] evaluation tool. Costa MT evaluation tool allows evaluating sentences by human mind more closely. Costa MT evaluation tool evaluate sentences in terms of 2 common aspects in Table 3.4.2. We have prepared a mix set of bilingual sentence pairs to present foreign language school instructor. Then, we got feedback of human evaluation scores.

CHAPTER 4

EVALUATION ANALYSIS

4.1 Automatic Evaluation Training Test

In this first section, 190 sentence pairs over 245 were tested to train system and rest of them is 55 sentence pairs were used for verify test by different automatic evaluation methods. Three selected MT services' such as Google, Bing and Yandex results were tested to train our bilingual sample set. And confidence interval and rate (95%) details according to declaration of 3.4.3. All similarity rates of texts are coming from automatic tools.

4.2 Automatic Metric Evaluation of Google

For this evaluation process, data consist of Turkish reference and Google translated sentences. Totally, 190 single, unique reference sentences are used to measure similarity rates with candidate sentence coming from Google translation service.

4.2.1 Evaluation Train Test of Google Service from English to Turkish

This evaluation part is made for Bleu+ metric especially because similarity scores of candidate and reference text can obtain one by one. It takes so many hours to get similarity scores for a big corpus. So, totally we examined over 190 sentence pairs to take similarity scores easily. Bilingual text pair test case consist of Google translated candidate text and reference texts within 50 simple, 50 complex, 50 compound and 40 complex-compound sentence structure.

The table on the following shows that distribution number of sentence in terms of their structure and average similarity scores of them in many various evaluation types:

Table 4.2.1.1: Average Scores of Evaluation Rates for All Structures of Turkish Sentence on Google

#	Sentence Structure	P	R	T	W	M	B	B+
50	**Simple**	35.8	37.6	15.4	13.7	15.5	13.6	47.9
50	**Complex**	42.0	42.7	26	22.6	19.2	15.7	36.5
50	**Compound**	47.2	48.7	31.0	26.8	20.9	14.8	36.8
40	**Compound-Complex**	44.3	46.4	28.8	26.5	19.0	12.5	39.7

These results are coming from examination of evaluation tests between Google translation service's outputs from English to Turkish language and already translated reference sentences by human experts which are about 50 simple, 50 Complex, 50 Compound and 40 Complex-Compound Turkish sentence pairs. Scores are between in the range of 0 and 100.

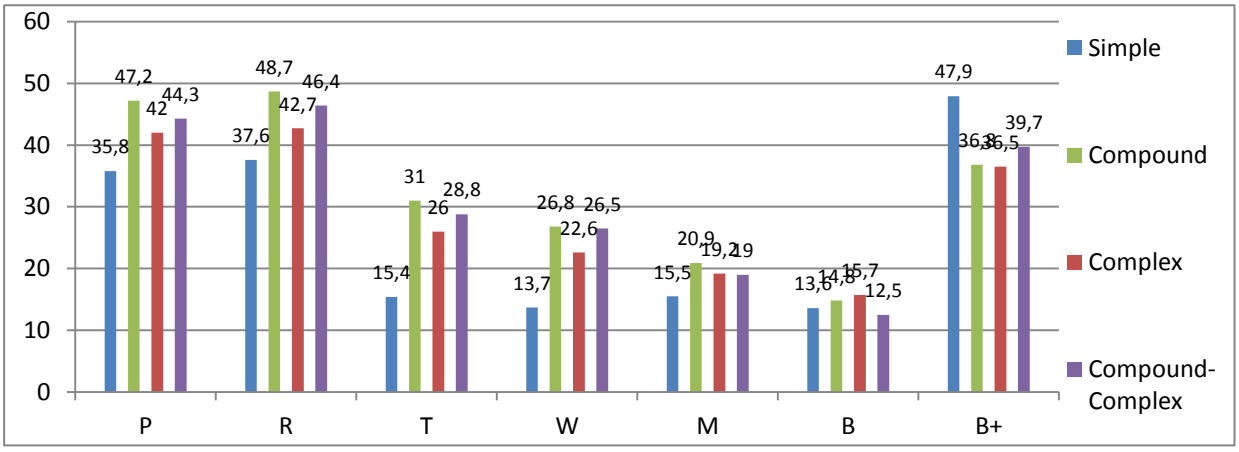


Figure 4.2.1.1: Average Scores of Evaluation Rates for All Structures of Turkish Sentence on Google

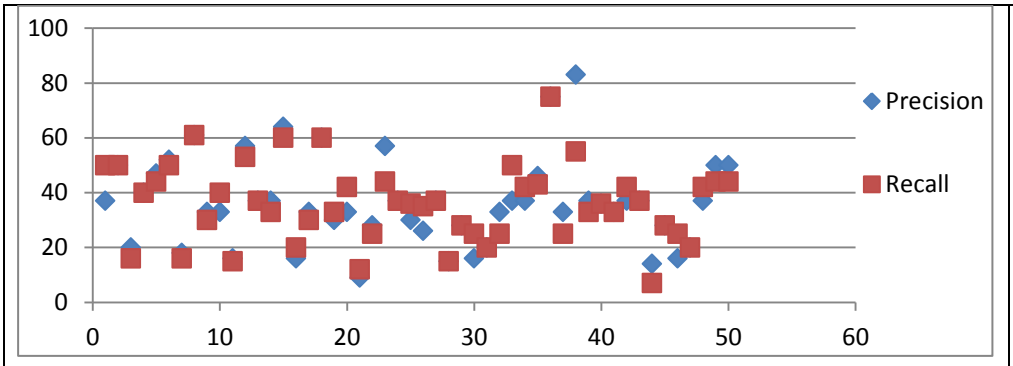
Bleu+ Rates has been obtained together all of sentence by bleu+ evaluation of MT tool. Together getting scores reflects different rates. But one by one scores more reliable. We decide to take 50 sentences to training test and 15 samples to verify test respectively. Excluding complex-compound sentence that for train 40 and 15 sentences to verify test. Scores are given by 95% confidence interval. Services are evaluated individually by metrics. The details are in table on the following:

Table 4.2.1.2: A Sample Sentence Similarity Scores of Google for Turkish Sentence Structures 1/2

Sent. Struc.	#	Source-En	Reference-Tr	Candidate Translation Tr	P	R	T	W	M	B	B+
Simple	**1**	China's Han Dynasty marked an official recognition of Confucianism.	Çin'in Han Hanedanı Konfüçyüscülüğü resmen tanıdı.	Çin'in Han Hanedanı Konfüçyüsçülük bir resmi tanıma işaretlenmiş.	37.0	50.0	17.0	17.0	24.7	19.1	36.0
	2	America's best known novelists, journalists, and editors attended a conference in New York last week.	Amerika'nın en ünlü romancıları, gazetecileri ve editörleri geçen hafta New York'ta bir konferansa katıldılar.	Amerika'nın en tanınmış romancılar, gazeteciler, editörler ve geçen hafta New York'ta bir konferansa katıldı.	53.3	61.5	46.2	46.2	29.4	25.3	31.0
	3	They decided to get married next month.	Gelecek ay evlenmeye karar verdiler.	Gelecek ay evlenmeye karar.	50.0	60.0	40.0	40.0	30.2	26.3	56.0
	…	…	…	…							
	50	Susan sang a solo and accompanied herself on the piano.	Susan solo bir parçayla piyanoda kendi kendine eşlik etti.	Susan solo seslendirdi ve piyano kendini eşlik etti.	50.0	44.4	44.4	44.4	21.6	20.0	54.0
Complex	**51**	The man known as “The Bulldozer” in Israel was “The Butcher” among its enemies.	İsrail'de "Buldozer" olarak bilinen Sharon, düşmanları tarafından "Kasap" olarak adlandırılıyordu.	İsrail'de "Bulldozer" olarak bilinen adam onun düşmanları arasında "Kasap" oldu.	50.0	50.0	40.0	40.0	21.3	18.5	39.0
	52	I wrote Jane a letter while she was away at camp.	Jane'e, o kamptayken bir mektup yazdım.	O kampta iken ben Jane'e mektup yazdı.	28.6	33.3	0.0	0.0	6.5	7.3	61.0
	53	The term "Stone Age" is used to describe a period of human evolution where stone was used as the hardest material for making tools.	"Taş Devri" terimi, insanlığın gelişim evrelerinden, aletlerin yapımında taşın en sert materyal olarak kullanıldığı dönemi tanımlamak için kullanılır.	Terimi "Taş Devri" taş alet yapımında en zor malzemesi olarak kullanılan insan evriminin bir dönemi tanımlamak için kullanılır.	50.0	50.0	27.8	27.8	23.3	27.8	17.0
	…	…	…	…							
	100	Numerous studies have shown that when smokers quit smoking, they sleep better in spite of temporary symptoms such as restlessness, anxiety and headache, which can persist for about ten days.	Pek çok çalışma; sigara içenlerin sigara içmeyi bıraktıklarında yaklaşık olarak on gün sürebilecek huzursuzluk, endişe ve baş ağrısı gibi geçici belirtilere rağmen daha iyi uyuduklarını göstermiştir.	Çeşitli çalışmalar tiryakiler sigarayı bırakma, onlar yaklaşık on gün boyunca sürebilen, huzursuzluk, anksiyete ve baş ağrısı gibi geçici semptomlar, rağmen iyi uyku olduğunu göstermiştir.	50.0	46.2	38.5	38.5	20.9	17.9	17.0

Table 4.2.1.3: A Sample Sentence Similarity Scores of Google for Turkish Sentence Structure 2/2

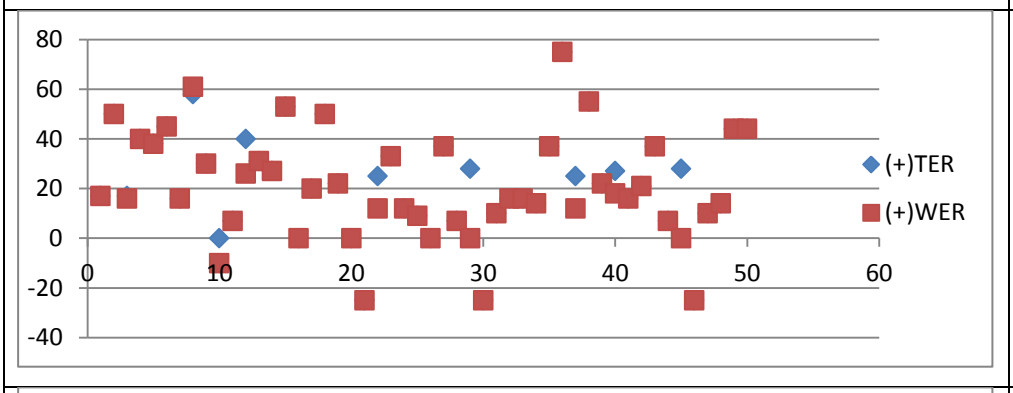
Sent. Struc.	#	Source-En	Reference-Tr	Candidate Translation Tr	P	R	T	W	M	B	B+
Compound	**101**	The classes ended early, but nobody in my class went home early.	Dersler erkenden bitti, ama sınıfımdan hiç kimse erkenden evine gitmedi.	Sınıflar erken bitti, ama benim sınıfta kimse eve erken gittim.	33.3	36.4	27.3	27.3	13.0	13.6	38.0
	102	His stories are always very long and boring, but we always listen to be polite.	Onun hikâyeleri her zaman uzun ve sıkıcıdır yine de biz nezaketimizden her zaman dinleriz.	Onun hikâyeleri her zaman çok uzun ve sıkıcı, ama biz her zaman kibar olmaya dinleyin.	47.1	61.5	23.1	23.1	25.1	8.5	56.0
	103	The boys walked down the road and their parents waved from the house.	Çocuklar yoldan aşağı yürüdüler ve anne babaları onlara evden el salladılar.	Çocuklar yolda yürüdü ve velileri evden salladı.	37.5	30.0	30.0	30.0	10.3	5.7	54.0
	…	…	…	…							
	150	Tom isn't always late, but he often is.	Tom her zaman geç kalmaz fakat sık sık kalır.	Tom her zaman geç değildir, ama o genellikle.	42.9	40.0	33.3	33.3	16.8	8.3	38.0
Compound-Complex	**151**	We had heard the assignment, but we did not understand it because the directions were confusing.	Görevi duymuştuk, ancak talimatlar kafa karıştırıcı olduğundan ne olduğunu anlayamadık.	Biz atama duymuştum, ama yön kafa karıştırıcı, çünkü biz onu anlamadı.	28.0	33.3	16.7	16.7	10.6	6.8	54.0
	152	Tom said he wasn't interested in Mary, but he seemed to always be looking towards the side of the room where she was.	Tom Mary ile ilgilenmediğini söyledi fakat o her zaman onun bulunduğu odanın tarafına doğru bakıyor gibi görünüyordu.	Tom Mary ilgi değildi dedi, ama o hep o oldu oda tarafına doğru bakıyor gibiydi.	41.0	38.9	33.3	33.3	17.8	11.0	24.0
	153	It is not the strongest of the species that survive, not the most intelligent, but the one most responsive to change.	O, yaşayan türlerin en güçlüsü değil, en zekisi değil fakat değişmek için en duyarlı olanıdır.	Bu, en zeki değil hayatta türlerin güçlü değil, ama değiştirmek için en duyarlı biri.	58.0	55.6	44.4	38.9	22.3	12.9	24.0
	…	…	…	…							
	190	Your English is grammatically correct, but sometimes what you say just doesn't sound like what a native speaker would say.	İngilizcen dilbilgisi bakımından doğru fakat bazen söylediğin tam olarak bir yerlinin söylediğine benzemiyor.	Sizin İngilizce dilbilgisi açısından doğru olduğunu, ancak bazen sadece anadili ne derdi benzemiyor ne demek.	55.0	57.9	21.1	15.8	20.1	5.7	81.0



	Prec.	Recall
Avr.	36.2	37.9
Std. Dev.	16.9	16.0
Conf. Intv.	36.2 ±4.	37.9 ±4.4

P: 31.5 < μx < 40.9

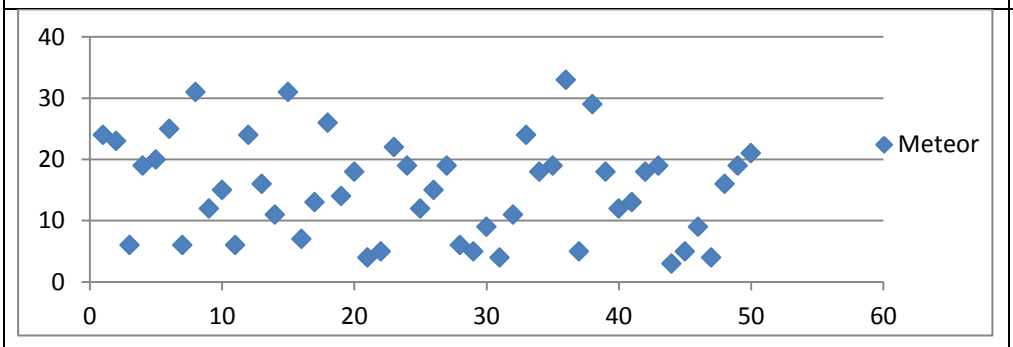
R: 33.5 < μx < 42.3



	Ter	Wer
Avr.	15.7	14.0
Std. Dev.	31.0	30.7
Conf. Intv.	15.7 ±8.6	14.0 ±8.5

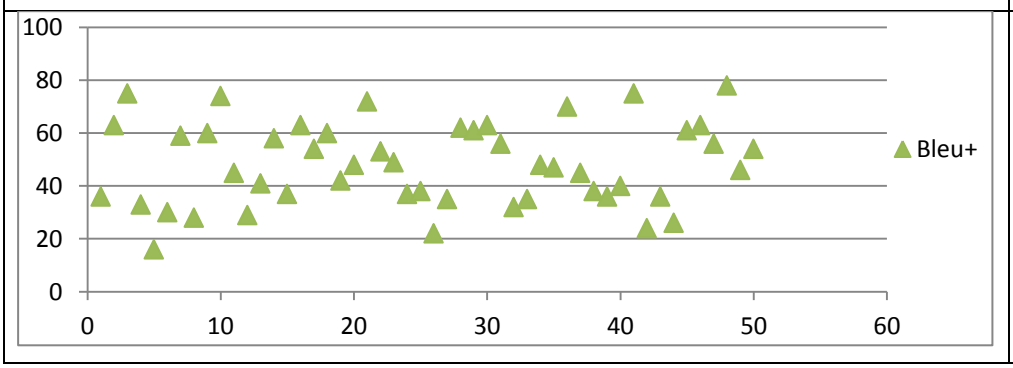
T: 7.1 < μx < 24.3

W: 5.5 < μx < 22.5



	Meteor
Avr.	16.0
Std. Dev.	9.0
Conf. Intv.	16.0 ± 2.

13.5< μx < 18.5



	Bleu+
Avr.	47.9
Std. Dev.	13.8
Conf. Intv.	47.9 ±3.8

44.1< μx < 51.8

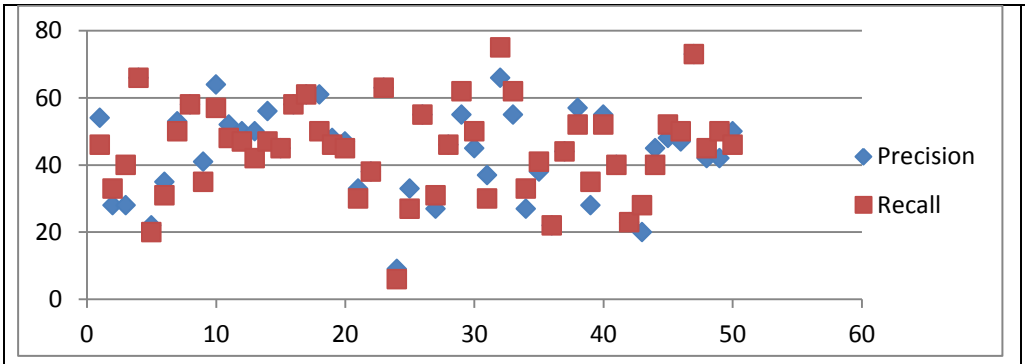
Figure 4.2.1.2: Average Scores of Evaluation Rates for Turkish Simple Sentences on Google

The table above includes 4 different figures and 4 mini tables which have explanation about related figure. Google translation service's outputs in Turkish language especially 50 simple sentences for system of train are evaluated with 6 different popular automatic evaluation methods. Results demonstrate that precision and recall methods are almost give same results also TER and WER, too. Meteor and Bleu+ mean scores are highly different.

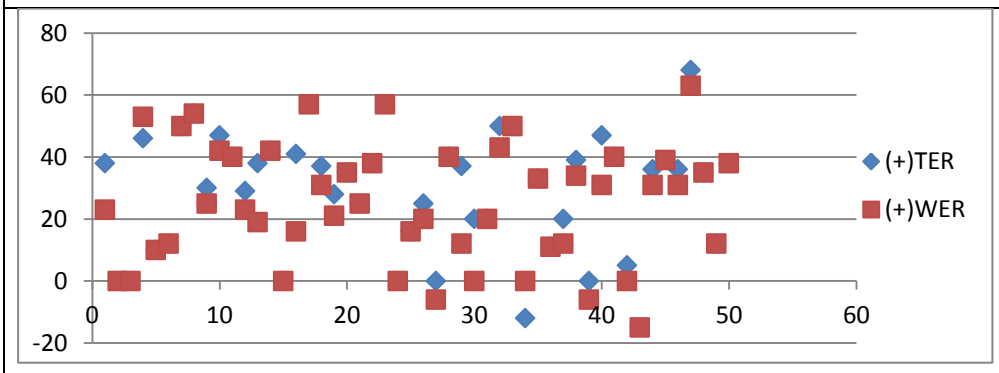
While Meteor methods give us 16 averages score over 100 and it has standard deviation of 9. So range of distribution is from 7 to 25 with standard deviation and in range of 13.5 and 18.5 with 95% confidence interval. However Bleu+ score gives us 48 average evaluation score over 50 individual sentence evaluation scores. Standard deviation is 15 so it has range from 33 to 63.

In general approach, the minimum standard deviation, 8, is at Meteor. Moreover, maximum standard deviation, 20, is coming from WER.

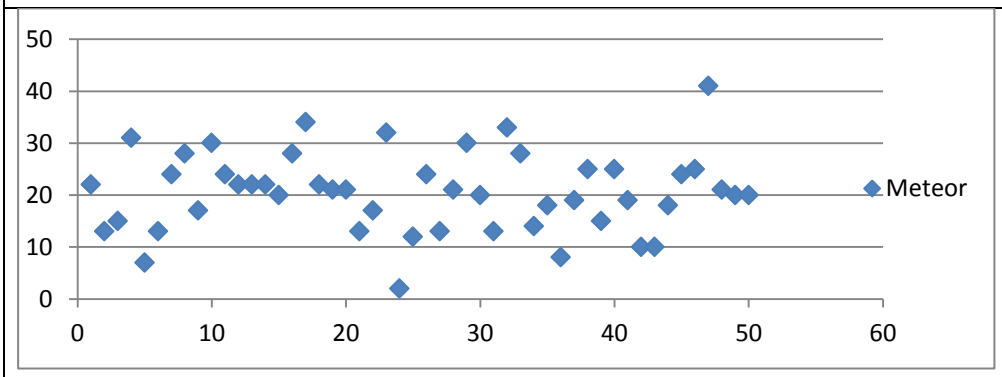
In terms of given simple reference sentences and metric measurement, Google translation service is giving many irrelevant words according to reference translation in simple sentence structure. So Recall rates are higher than Precision rates naturally.



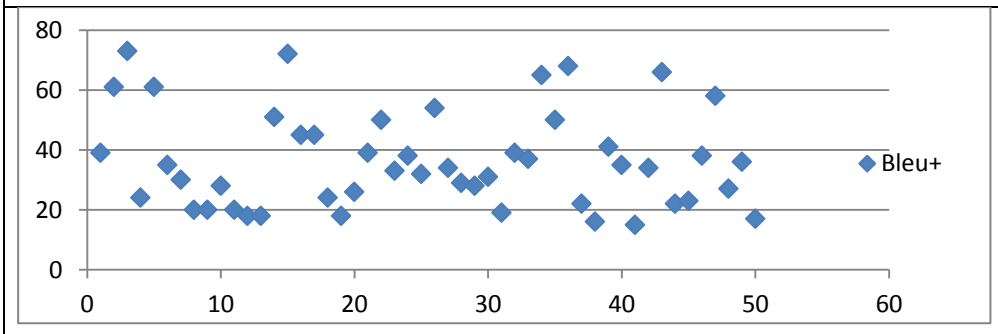
	Prec.	Recall
Avr.	42.4	43.1
Std. Dev.	±14.9	±14.7
Conf. Intv.	±4.1	±4.1
Precision	38.3< μx <46.6	
Recall	39.0< μx <47.1	



	Ter	Wer
Avr.	26.3	22.9
Std. Dev.	20.7	21.4
Conf. Intv.	5.7	5.9
TER	20.5<μx<32.0	
WER	17.0<μx<28.8	



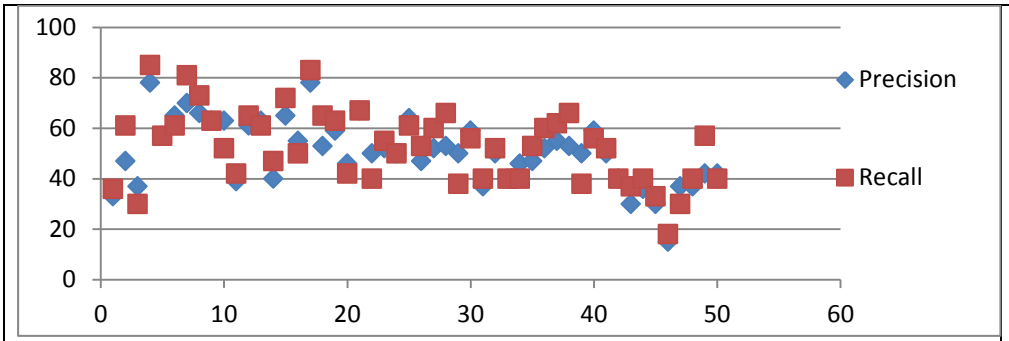
	Meteor
Avr.	19.7
Std. Dev.	8.2
Conf. Intv.	2.3
M	17.4< μx <22.0



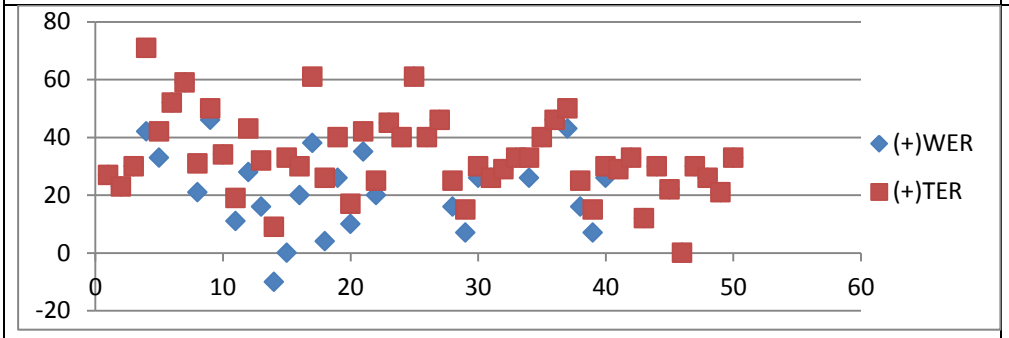
	Bleu+
Avr.	36.5
Std. Dev.	15.2
Conf. Intv.	4.2
	32.3< μx <40.8

Figure 4.2.1.3: Average Scores of Evaluation Rates for Turkish Complex Sentences on Google

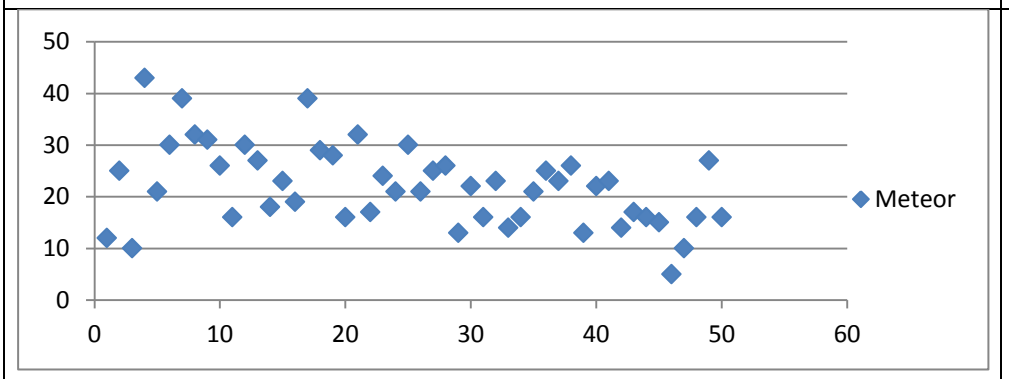
It is seen clearly that more densely explanations are occurred in Turkish reference translation over complex structure, again. And in terms of comparatively examination of Recall and Bleu+, complex sentence translation process is more successfully than simple ones.



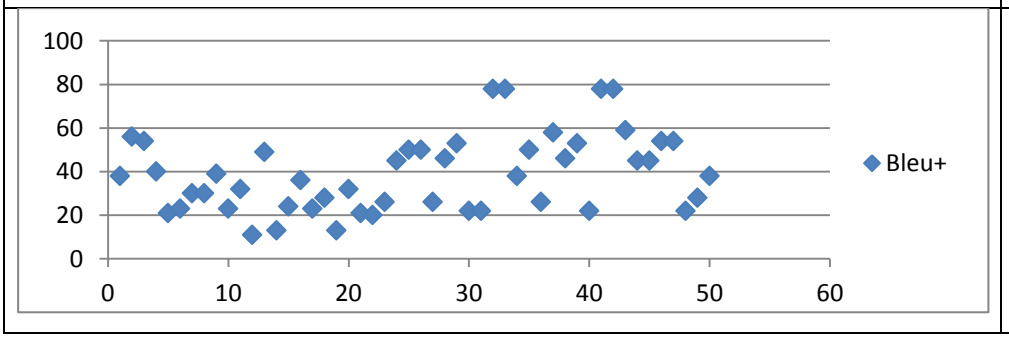
	Prec.	Recall
Avr.	50.9	52.9
Std. Dev.	12.8	14.6
Conf. Intv.	3.6	4.1
P	47.4< μx <54.5	
R	48.9< μx <57.0	



	Ter	Wer
Avr.	33.5	27.8
Std. Dev.	14.0	15.1
Conf. Intv.	3.9	4.2
Ter Range	29.6< μx <37.4	
Wer Range	23.7< μx <32.0	



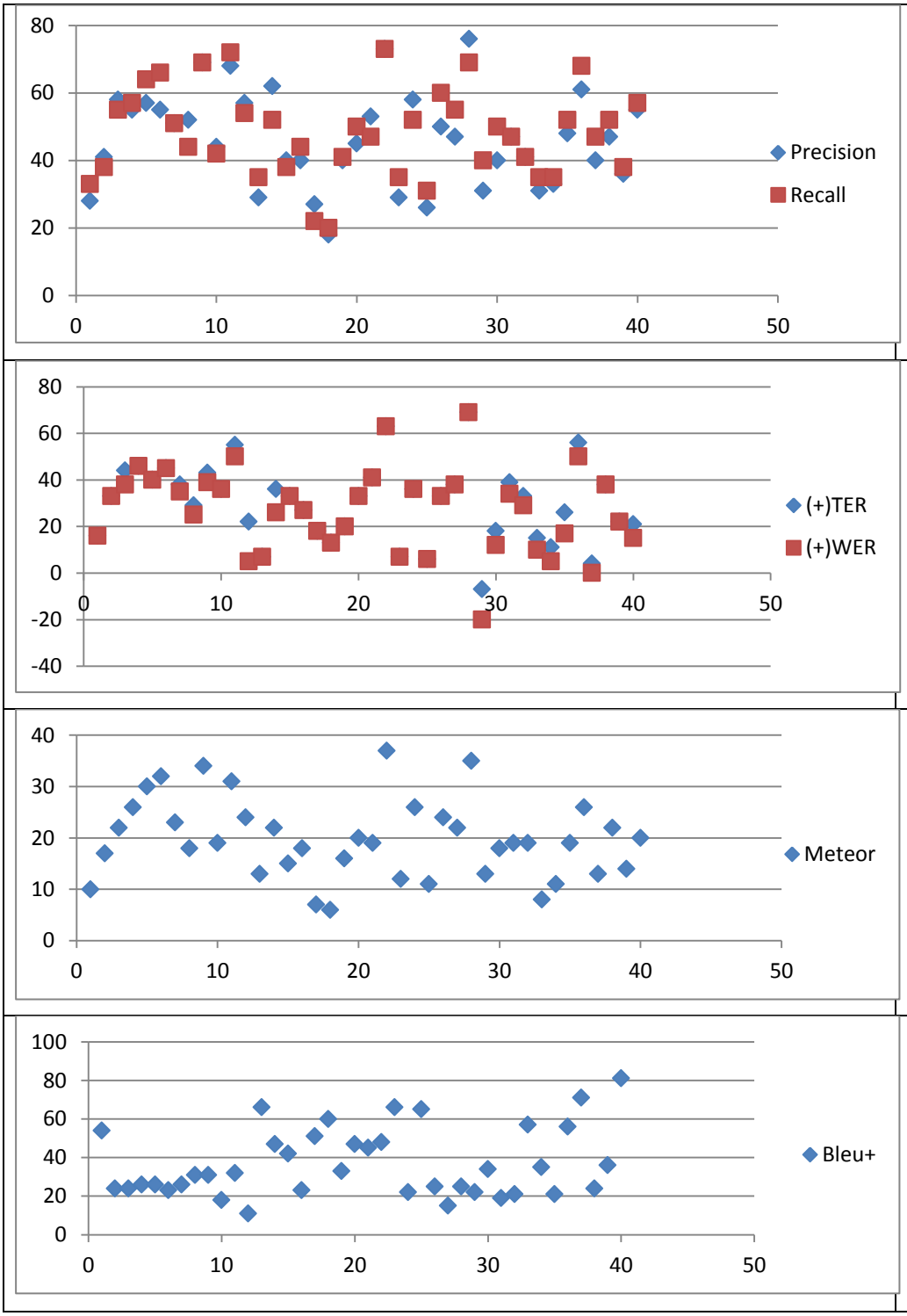
	Meteor
Avr.	22.7
Std. Dev.	7.9
Conf. Intv.	2.2
Range	20.5< μx <24.9



	Bleu+
Avr.	38.9
Std. Dev.	17.6
Conf. Intv.	4.9
Range	34.0< μx <43.8

Figure 4.2.1.4: Average Scores of Evaluation Rates for Turkish Compound Sentences on Google

I detect that generally confidence interval bounds are 20% or 25% of standard deviation. So we may assume at which rate general standard deviation estimation. On compound Sentence alignment of words quality is higher than previous ones.



	P	R
Avr.	46.8	48.7
Std. Dev.	13.8	13.2
Conf. Intv.	4.3	4.1
Prec. Range	42.5< µx <51.1	
Recall Range	44.6< µx <52.8	

	T	**W**
Avr.	30.5	27.7
Std. Dev.	16.6	17.9
Conf. Intv.	5.2	5.5
TER Range	25.4< µx <35.7	
WER Range	22.1< µx <33.2	

	M
Avr.	20.2
Std. Dev.	7.7
Conf. Intv.	2.4
Range	17.9< µx <22.6

	B+
Avr.	37.1
Std. Dev.	17.7
Conf. Intv.	5.5
Range	31.6< µx <42.6

Figure 4.2.1.5: Average Scores of Evaluation Rates for Turkish Complex-Compound Sentences on Google

It can be said that averages of scores are going to a stable balanced value whether each score differ from one another. Also the automatic method names are abbreviated with their head character like Blue and Precision, etc. As well as Recall and Precision score rates are compatible likely some other metrics, TER and WER, show that almost same values.

4.2.2 Evaluation Train Test of Google Service from Turkish to English

Especially Bleu+ comparison tests are made also for in translation from Turkish to English over 190 sentences.

It is seen that for English sentence evaluation it can be talked about the results that precession rates/scores either equal or bigger than recall score rates over entire corpus. The table on following that shows average score of sentence structure by auto metrics:

Table 4.2.2.1: Average Scores of Evaluation Rates for All Structures of English Sentences on Google

#		Precision	Recall	TER	WER	Meteor	Bleu	Bleu+
50	**Simple**	59.8	57.2	42.9	38.2	28.9	31.9	45.5
50	**Complex**	57.1	56.7	34.1	23.8	24.8	24.8	29.4
50	**Compound**	57.4	56.1	38.1	26.4	25.3	25.6	31.4
40	**Compound-Complex**	56.3	53.1	35.4	28.5	23.5	23.9	27.5

Result shows that in different sentence structures, Google may translate Turkish source sentence to English language variously as seen in the following Figure 4.2.2.1:

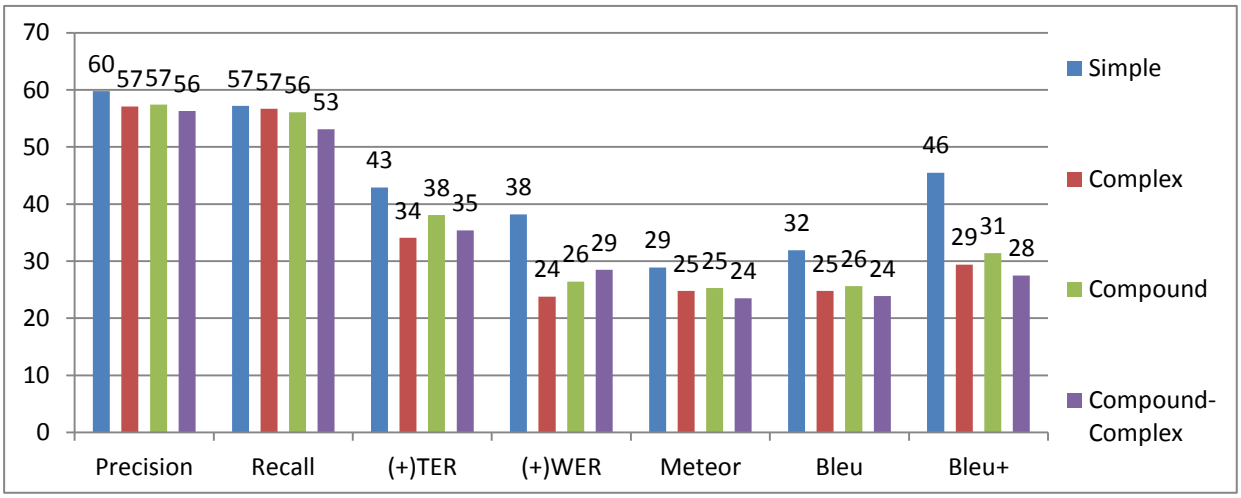


Figure 4.2.2.1: Average Scores of Evaluation Rates for All Structures of English Sentences on Google

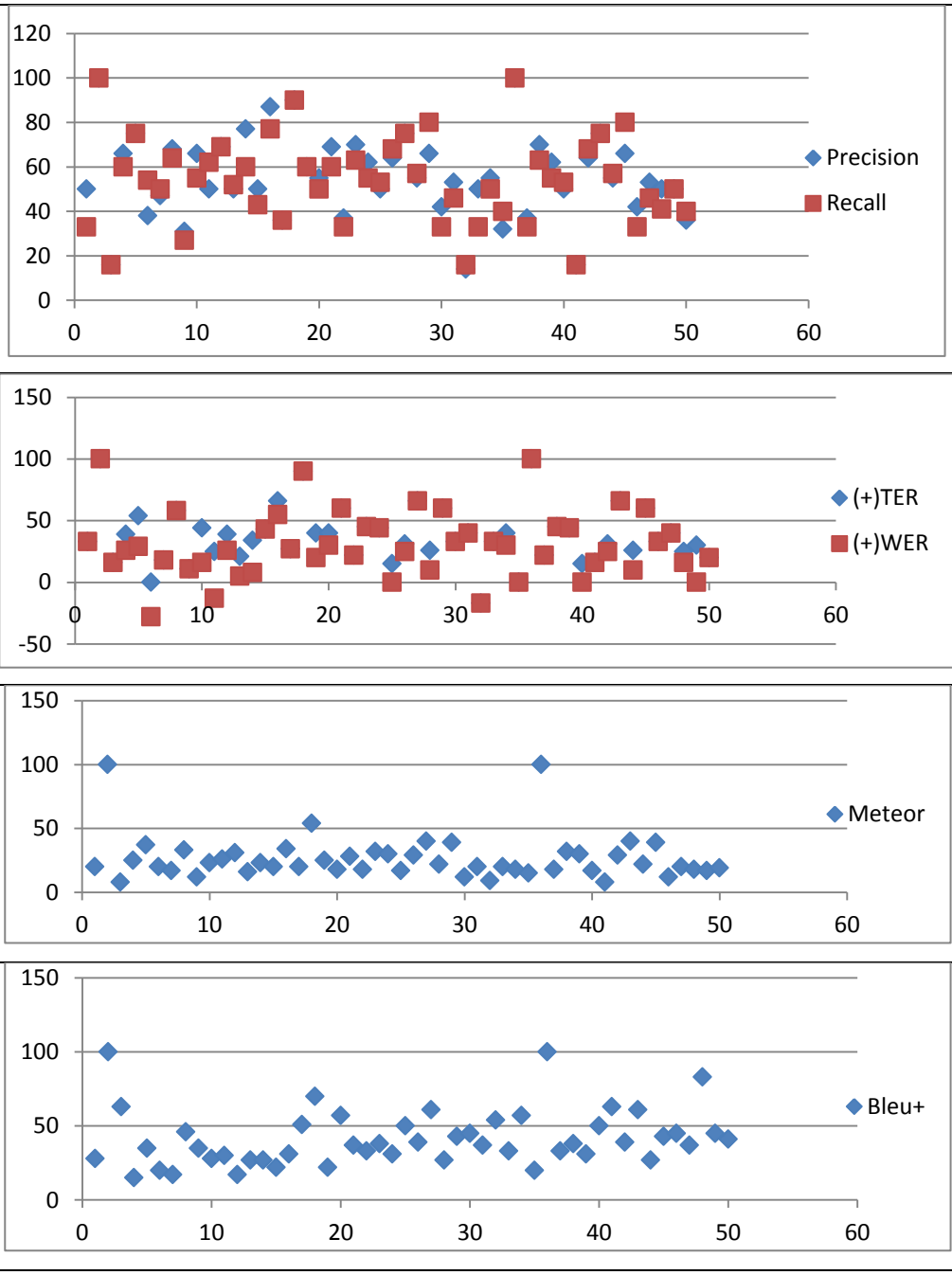
So it is seen clearly again that different sentences structures can be translated in different quality. Precision and recall rates almost same and complex and compound-complex sentences are almost in same level. But compound sentences are different a little bit. Also simple sentences are obviously separated with high scores.

Table 4.2.2.2: Sentence Similarity Scores between English Google Candidate and Reference Sentences 1/2

Sent. Struc.	#	Source-Tr	Reference-En	Google En	P	R	T	W	M	B	B+
Simple	1	**Çin'in Han Hanedanı Konfiçyüsçülüğü resmen tanıdı.**	**China's Han Dynasty marked an official recognition of Confucianism.**	**China's Han Dynasty Confucianism officially recognized.**	33.3	33.3	33.3	21.0	18.1	28.0	33.3
	2	**Amerika'nın en ünlü romancıları, gazetecileri ve editörleri geçen hafta New York'ta bir konferansa katıldılar.**	**America's best known novelists, journalists, and editors attended a conference in New York last week.**	**America's most famous novelists, journalists and editörleri geç attended a conference in New York last week.**	73.3	66.7	66.7	33.1	55.0	47.0	73.3
	3	**Gelecek ay evlenmeye karar verdiler.**	**They decided to get married next month.**	**They decided to get married next month.**	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	...	**…**	**…**	**…**							
	50	**Susan solo bir parçayla piyanoda kendi kendine eşlik etti.**	**Susan sang a solo and accompanied herself on the piano.**	**Susan is a solo piano piece was accompanied by a self.**	36.4	40.0	20.0	20.0	19.6	8.9	41.0
Complex	51	**İsrail'de "Buldozer" olarak bilinen Sharon, düşmanları tarafından "Kasap" olarak adlandırılıyordu.**	**The man known as “The Bulldozer” in Israel was “The Butcher” among its enemies.**	**In Israel, the "Bulldozer" Sharon, known as enemies by the "Butcher" was called.**	35.7	35.7	21.4	0.0	8.0	5.4	32.0
	52	**Jane'e, o kamptayken bir mektup yazdım.**	**I wrote Jane a letter while she was away at camp.**	**Jane, she wrote a letter to the camp.**	62.5	45.5	36.4	27.3	21.4	11.2	37.0
	53	**"Taş Devri" terimi, insanlığın gelişim evrelerinden, aletlerin yapımında taşın en sert materyal olarak kullanıldığı dönemi tanımlamak için kullanılır.**	**The term "Stone Age" is used to describe a period of human evolution where stone was used as the hardest material for making tools.**	**The Flintstones, the term stages of the development of mankind, the hardest stone in the construction of the instruments used in this study is used to describe the period.**	37.9	44.0	12.0	-4.0	17.6	9.8	10.0
	…	…	…	…							
	100	**Pek çok çalışma; sigara içenlerin sigara içmeyi bıraktıklarında yaklaşık olarak on gün sürebilecek huzursuzluk, endişe ve baş ağrısı gibi geçici belirtilere rağmen daha iyi uyuduklarını göstermiştir.**	**Numerous studies have shown that when smokers quit smoking, they sleep better in spite of temporary symptoms such as restlessness, anxiety and headache, which can persist for about ten days.**	**Many studies; When smokers quit smoking, which can take approximately ten days restlessness, anxiety, and transient symptoms such as headache, sleep better showed that despite.**	64.0	53.3	33.3	13.3	24.5	18.6	29.0

Table 4.2.2.3: Sentence Similarity Scores between English Google Candidate and Reference Sentences 2/2

Sent. Struc.	#	Source-Tr	Reference-En	Candidate Translation- En	Prec	Recall	TER	WER	Meteor	Bleu	Bleu+
Compound	101	Dersler erkenden bitti, ama sınıfımdan hiç kimse erkenden evine gitmedi.	The classes ended early, but nobody in my class went home early.	Courses ran out early, but nobody from my class had to go home early.	50.0	58.3	41.7	41.7	31.0	24.7	32.0
	102	Onun hikayeleri herzaman uzun ve sıkıcıdır yinede biz nezaketimizden her zaman dinleriz.	His stories are always very long and boring, but we always listen to be polite.	His stories are always long and boring again, we always listen to our kindness are.	66.7	66.7	60.0	60.0	33.4	36.7	36.0
	103	Çocuklar yoldan aşağı yürüdüler anne-babaları onlara evden el salladılar.	The boys walked down the road and their parents waved from the house.	Parents of children marched down the road waving them at home.	36.4	30.8	23.1	15.4	13.4	11.5	24.0
	…	…	…	…							
	150	Ne üçüncü tarafların çıkarlarını baltalar ne de Çin'in uluslar arası yükümlülüklerini ihlal eder.	It doesn't undermine any other party's interests or violate China's international obligations.	What undermine the interests of third parties, nor would violate China's international obligations.	46.2	50.0	25.0	25.0	28.5	26.8	21.0
Compound-Complex	151	Görevi duymuştuk, ancak talimatlar kafa karıştırıcı olduğundan ne olduğunu anlayamadık.	We had heard the assignment, but we did not understand it because the directions were confusing.	We heard the task, but the instructions are confusing, I could not understand what happened.	46.7	43.8	31.3	25.0	16.3	8.5	49.0
	152	Tom Mary ile ilgilenmediğini söyledi fakat o her zaman onun bulunduğu odanın tarafına doğru bakıyor gibi görünüyordu.	Tom said he wasn't interested in Mary, but he seemed to always be looking towards the side of the room where she was.	Tom said that dealing with Mary, but she always looks toward the side of the room where it looked like.	60.0	52.2	43.5	43.5	25.8	27.3	26.0
	153	O, yaşayan türlerin en güçlüsü değil, en zekisi değil fakat değişmek için en duyarlı olanıdır.	It is not the strongest of the species that survive, not the most intelligent, but the one most responsive to change.	It is not the strongest of the species alive, not the most intelligent, but is most susceptible to change.	84.2	76.2	76.2	76.2	41.6	59.1	53.0
	…										
	190	Sezgileri gerçekten kuvvetli bir insan bütün bir durumu sadece birkaç ipucuyla çözebilir. Bu olmak istediğim kişi türüdür.	A really perceptive person can figure out a whole situation with just a few clues. That's the kind of person I want you to become.	Intuition is really a strong man can solve all the cases with only a few clues. This is the kind of person I want to be.	53.9	56.0	44.0	40.0	23.5	28.1	25.0



	P	R
Avr.	59.8	57.2
Std. Dev.	19.7	19.9
Conf. Intv.	5.5	5.5
Prec. Range	54.4< μx <65.3	
Recall Range	51.7< μx <62.7	

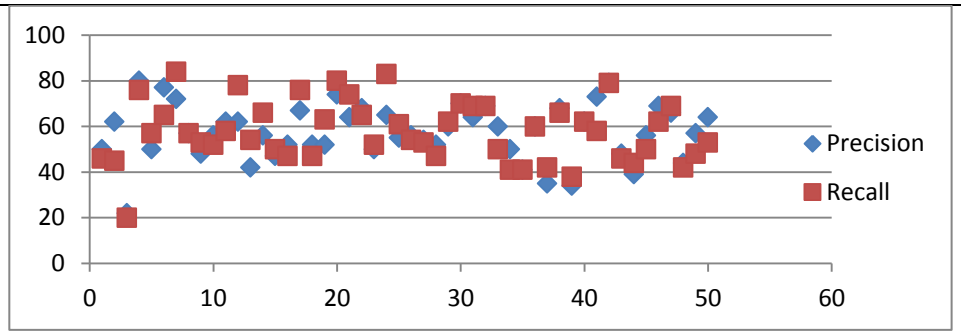
	T	W
Avr.	42.9	38.2
Std. Dev.	26.8	30.2
Conf. Intv.	7.4	8.4
TER Range	35.5< μx <50.3	
WER Range	29.8< μx <46.5	

	M
Avr.	28.9
Std. Dev.	18.0
Conf. Intv.	5.0
Range	23.9< μx <33.9

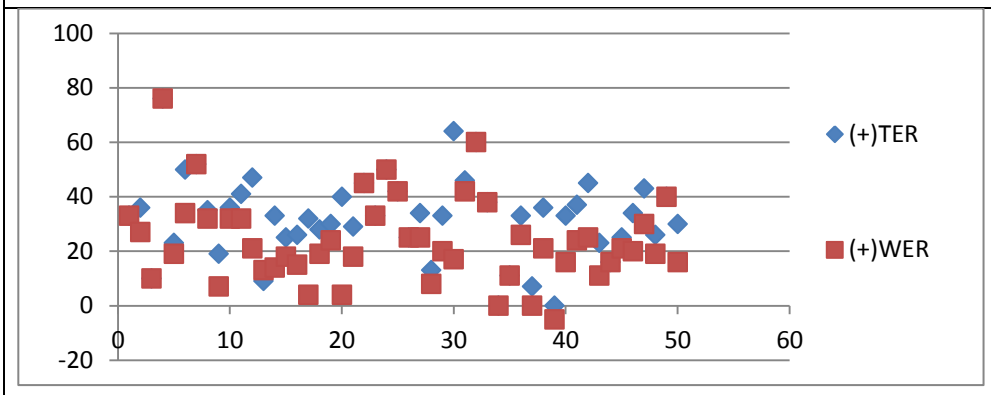
	B+
Avr.	45.5
Std. Dev.	17.3
Conf. Intv.	4.8
Range	40.7< μx <50.3

Figure 4.2.2.2: Average Scores of Evaluation Rates for English Simple Sentences on Google

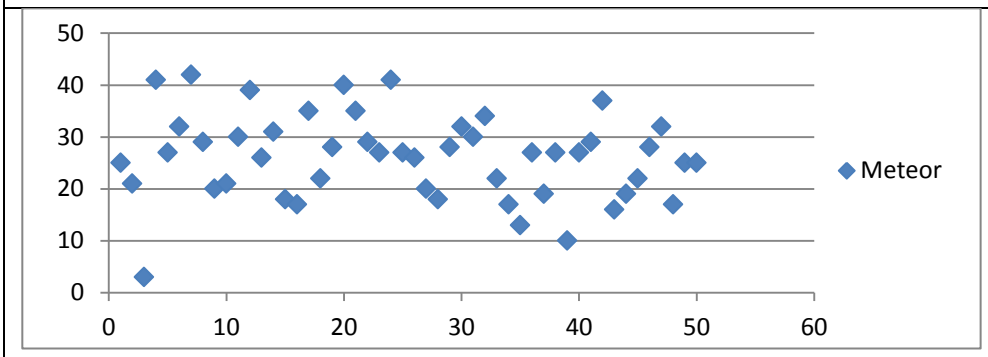
Simple sentence translation rates from Turkish to English are better than the translation from English to Turkish one. In addition, candidate sentence lengths are shorter than reference sentence length generally. Since sentence length so short some of translations are one-to-one the same.



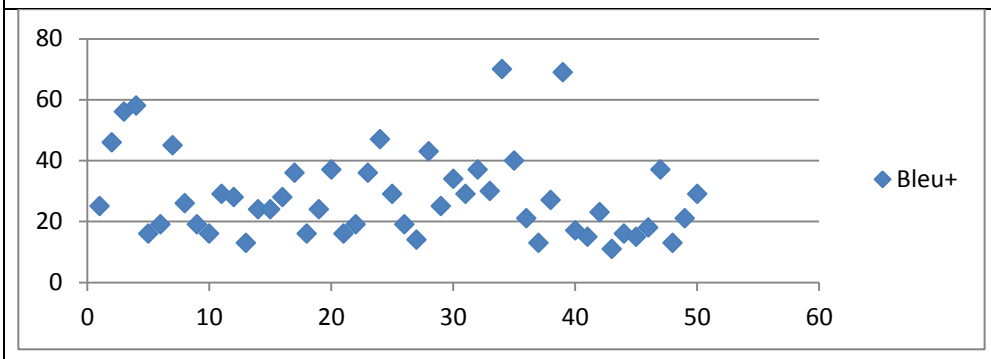
	P	R
Avr.	57.1	56.7
Std. Dev.	13.5	14.2
Conf. Intv.	3.7	3.9
Prec. Range	53.3< μx <60.8	
Recall Range	52.8< μx <60.7	



	TER	WER
Avr.	57.1	56.7
Std. Dev.	13.5	14.2
Conf. Intv.	3.7	3.9
Ter Range	53.3< μx <60.8	
Wer Range	52.8< μx <60.7	



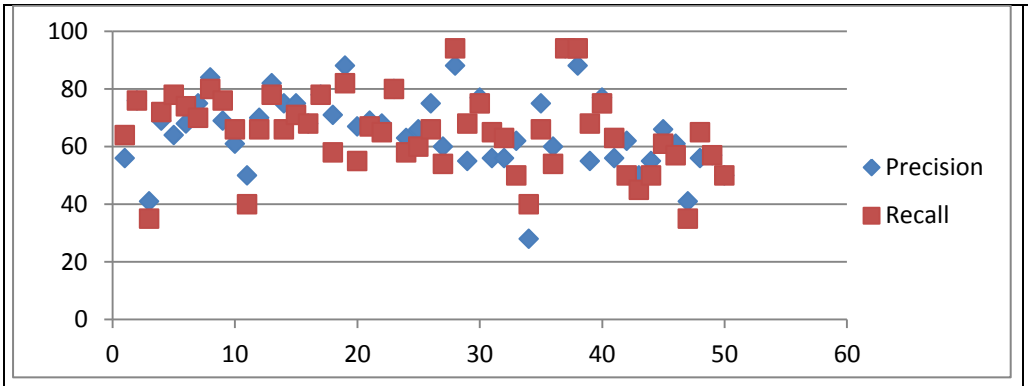
Meteor	
Avr.	24.8
Std. Dev.	8.9
Conf. Intv.	2.5
Range	22.4< μx <27.3



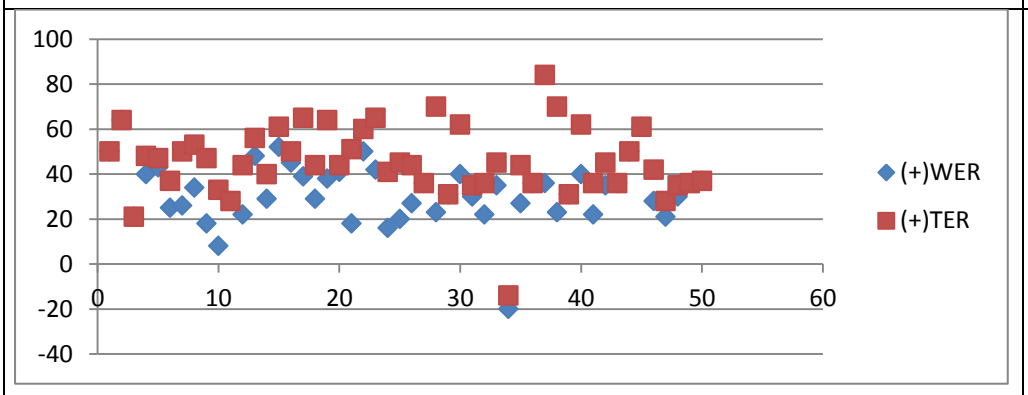
	Bleu+
Avr.	29.4
Std. Dev.	14.5
Conf. Intv.	4.0
Range	25.4< μx <33.5

Figure 4.2.2.3: Average Scores of Evaluation Rates for English Complex Sentences on Google

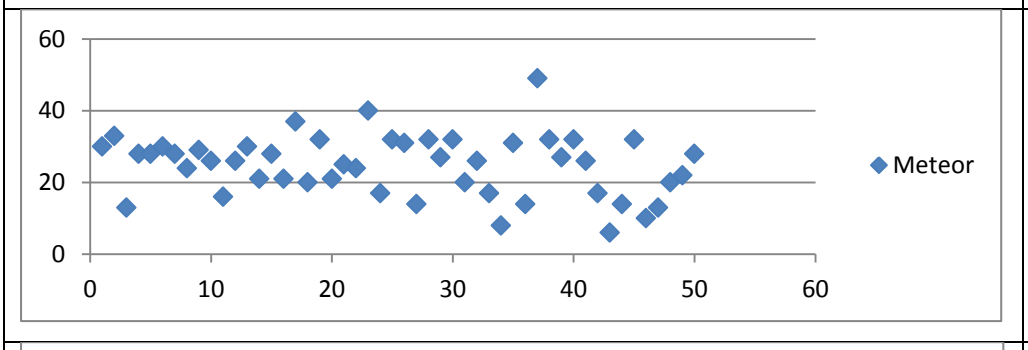
With longer sentence structure naturally, complex Turkish to English translations give us unaligned structure as well.



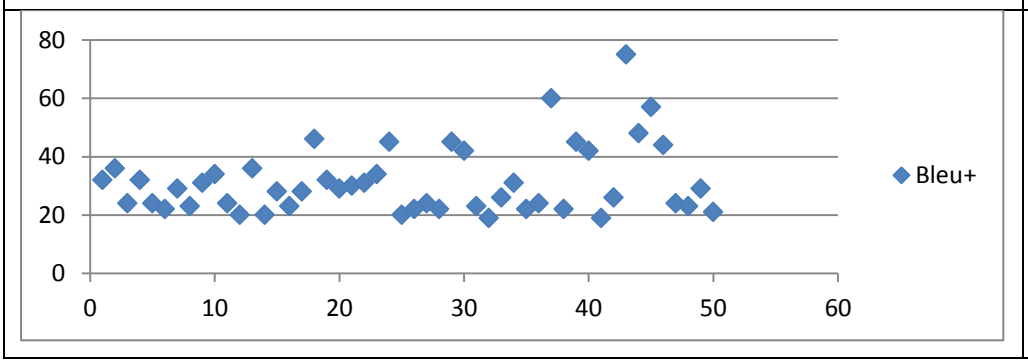
	Prec.	Recall
Avr.	57.4	56.1
Std. Dev.	13.5	13.8
Conf. Intv.	3.7	3.8
Prec. Range	53.6< µx <61.1	
Recall Range	52.2< µx <59.9	



	T	W
Avr.	38.1	26.4
Std. Dev.	15.4	13.8
Conf. Intv.	4.3	3.8
TER Range	33.8< µx <42.3	
WER Range	22.6< µx <30.3	



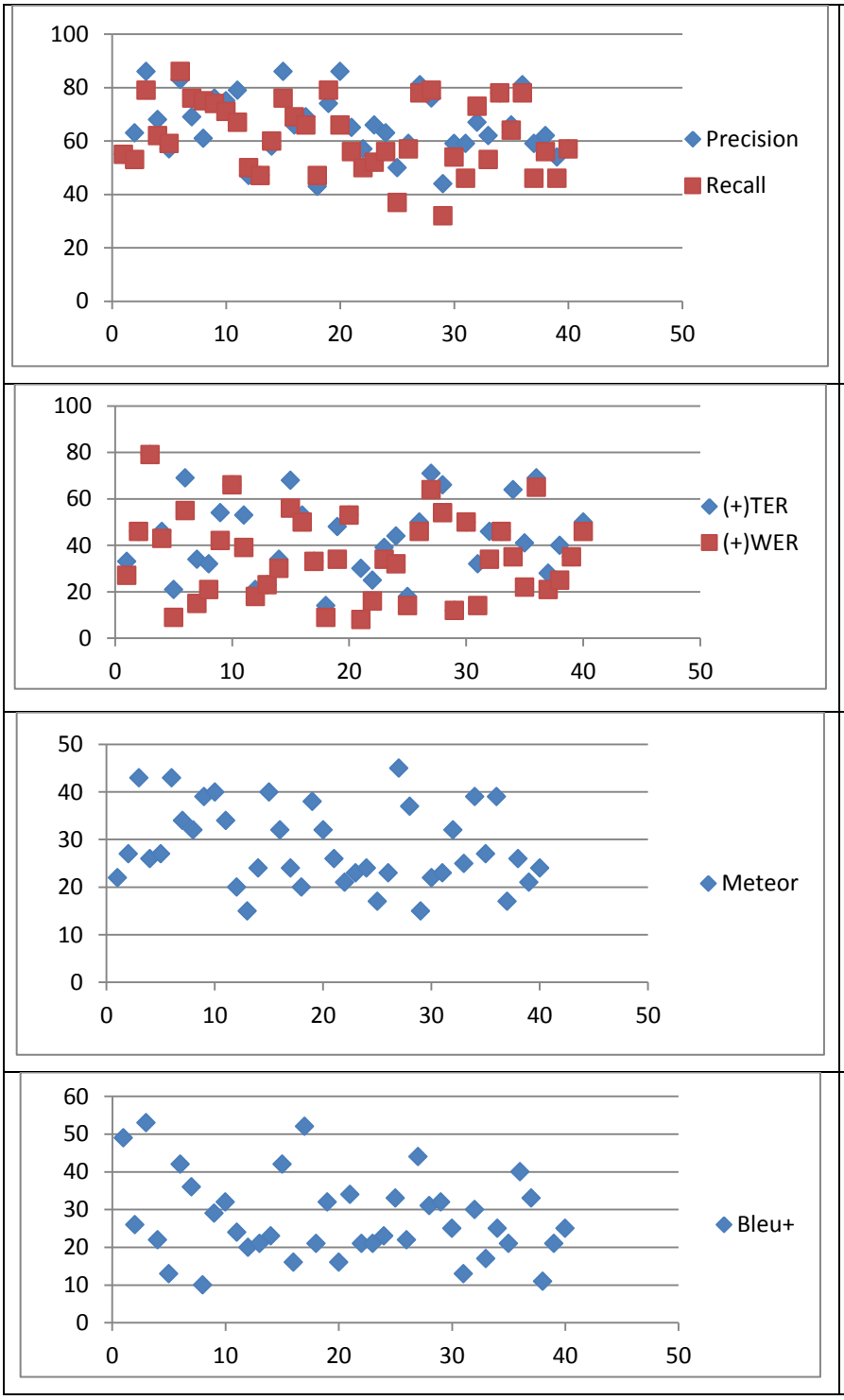
	Meteor
Avr.	25.3
Std. Dev.	8.5
Conf. Intv.	2.4
Range	22.9< µx <27.7



	Bleu+
Avr.	31.4
Std. Dev.	11.8
Conf. Intv.	3.3
Range	28.1< µx <34.6

Figure 4.2.2.4: Average Scores of Evaluation Rates for English Compound Sentences on Google

Since Compound sentence has more than one simple sentence, Precision and Recall rates are at high level but alignment of words over all sentence are not as well as at simple and complex sentences.



	Prec.	Recall
Avr.	56.3	53.1
Std. Dev.	13.4	13.1
Conf. Intv	4.2	4.1
Prec. Range	52.2< µx <60.5	
Recall Range	49.0< µx <57.1	

	T	W
Avr.	35.4	28.5
Std. Dev	16.2	17.8
Conf. Intv.	5.0	5.5
TER Range	30.4< µx <40.4	
WER Range	23.0< µx <34.1	

	Meteor
Avr.	23.5
Std. Dev.	8.0
Conf. Intv.	2.5
Range	21.0< µx <26.0

	Bleu+
Avr.	27.5
Std. Dev.	10.9
Conf. Intv.	3.4
Range	24.1< µx <30.9

Figure 4.2.2.5: Average Scores of Evaluation Rates for English Complex-Compound Sentences on Google

Although Precision and Recall rates are almost same, alignment of words are not better according to TER, WER, Bleu, Meteor and Bleu+ because of densely meaning.

4.3 Automatic Metric Evaluation of Bing

In this sub section, Bing translation service is evaluated in terms of language and metrics.

4.3.1 Evaluation Train Test of Bing Service from English to Turkish

The table on the following shows that 50 simple sentences, translated from English to Turkish languages, are compared with given reference sentence.

Table 4.3.1.1: Average Scores of Evaluation Rates for All Structures of Turkish Sentences on Bing

#	Sent. Str.	P	R	T	W	M	B	B+
50	**Simple**	43.3	43.6	27.1	27.0	19.9	17.2	64.5
50	**Complex**	50.3	47.8	32.7	25.7	19.1	10.3	48.4
50	**Compound**	48.9	48.2	33.7	29.2	19.9	13.1	41.8
40	**Complex-Compound**	46.7	46.0	29.8	26.1	18.2	10.5	34.5

Results prove that Bing translation service is better than Google in terms of translation output quality over simple sentences. And the other strong side of Bing is this that the translation from English to Turkish is closer than Google. The details are on the following table and figures.

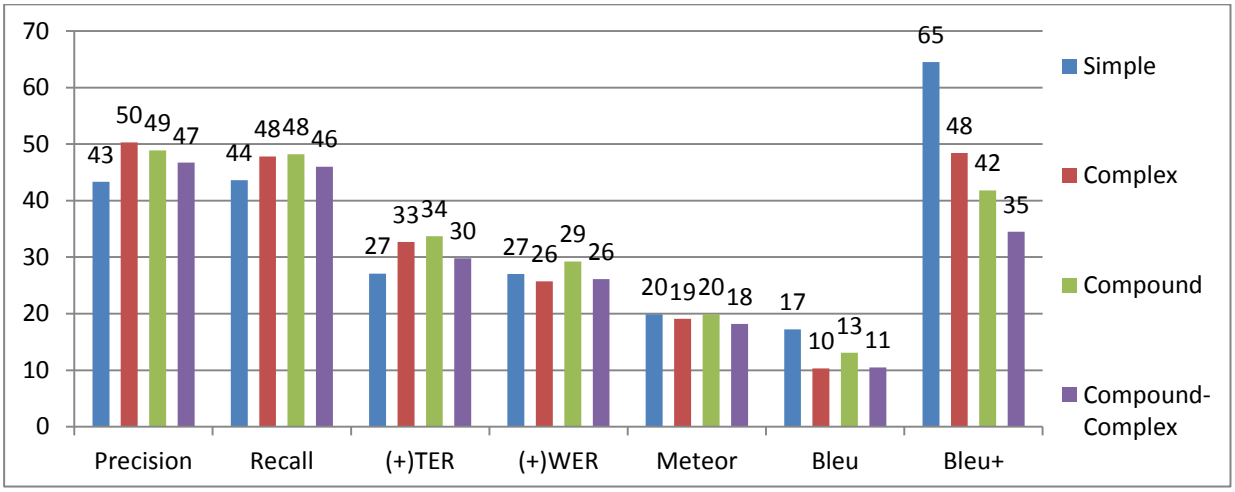
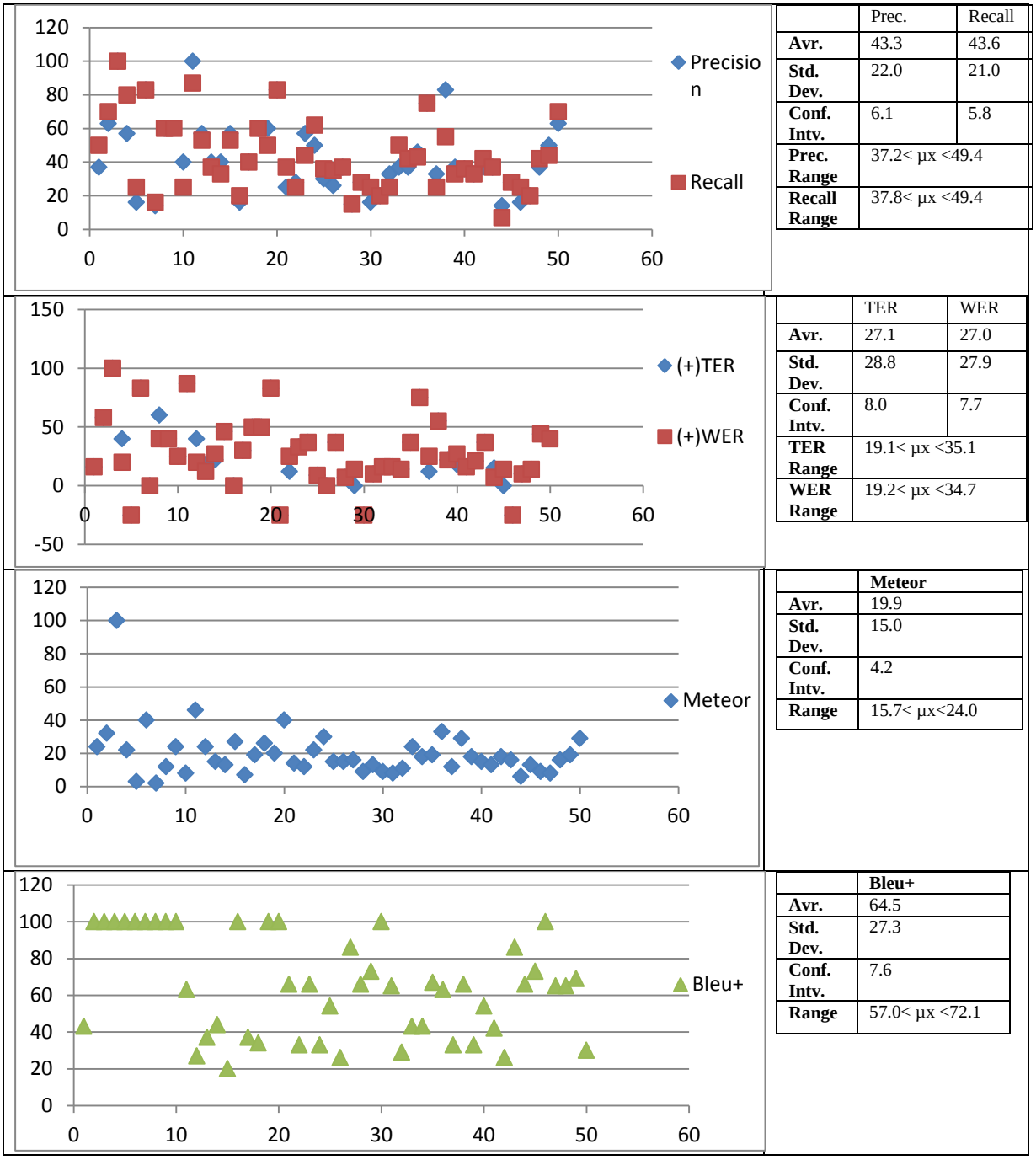


Figure 4.3.1.1: Average Scores of Evaluation Rates for All Structures of Turkish Sentences on Bing

It is very remarkable that however metrics give lower scores exclude Bleu+, it gives higher and the highest score especially on simply sentence structure at the end of similarity evaluation test. This is the reason that suffixes are very significant as a root. Because of

Turkish linguistic rules, some addition words are combined after word root. Bing translations service has high correlation about root of words but there is a little bit difference on suffix and word alignment level. But it is observed that performance of Bing getting decrease from complex to compound-complex sentence structure.



	Prec.	Recall
Avr.	43.3	43.6
Std. Dev.	22.0	21.0
Conf. Intv.	6.1	5.8
Prec. Range	37.2< µx <49.4	
Recall Range	37.8< µx <49.4	

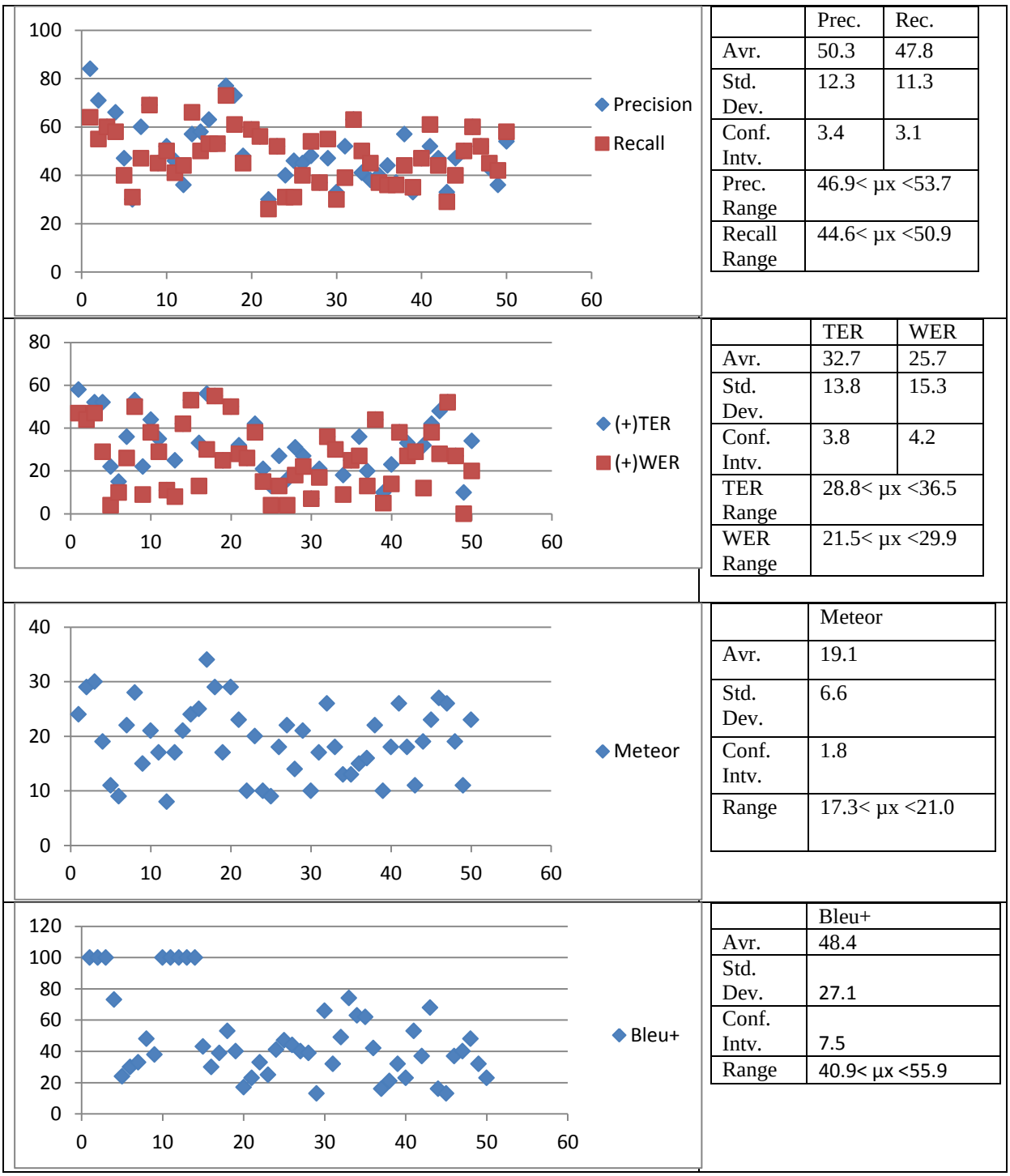
	TER	WER
Avr.	27.1	27.0
Std. Dev.	28.8	27.9
Conf. Intv.	8.0	7.7
TER Range	19.1< µx <35.1	
WER Range	19.2< µx <34.7	

	Meteor
Avr.	19.9
Std. Dev.	15.0
Conf. Intv.	4.2
Range	15.7< µx<24.0

	Bleu+
Avr.	64.5
Std. Dev.	27.3
Conf. Intv.	7.6
Range	57.0< µx <72.1

Figure 4.3.1.2: Average Scores of Evaluation Rates for Turkish Simple Sentences on Bing

Results are showed that there many same root but different suffixes on same root in the sentence. So suffix based measurement metric is Bleu+ can assess simple sentences translated by Bing better than other metrics and closer to human judgment. Individually test figures shows that corpus is consist of different type of sentence. So it is corresponding to general usage.



	Prec.	Rec.
Avr.	50.3	47.8
Std. Dev.	12.3	11.3
Conf. Intv.	3.4	3.1
Prec. Range	46.9< μx <53.7	
Recall Range	44.6< μx <50.9	

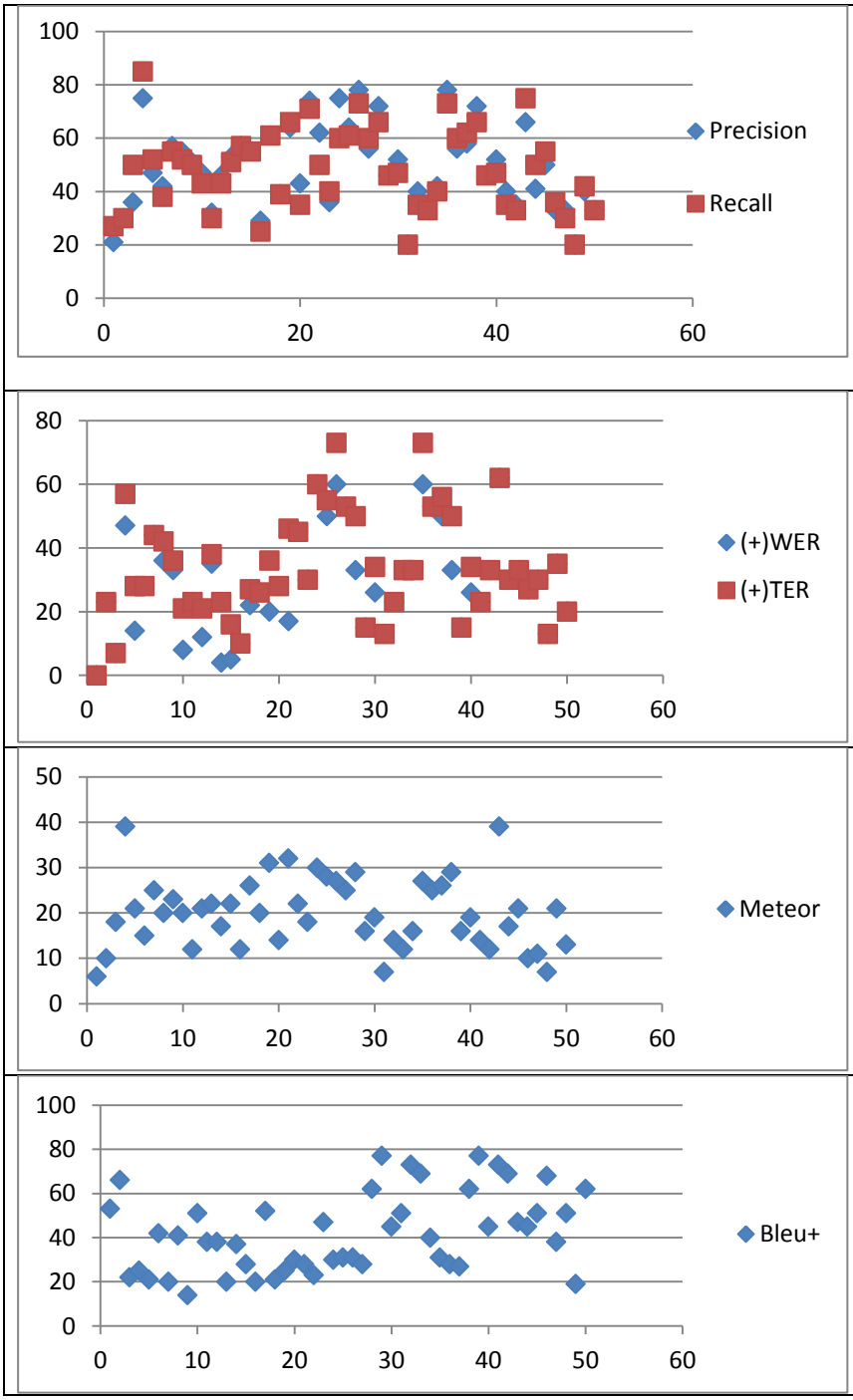
	TER	WER
Avr.	32.7	25.7
Std. Dev.	13.8	15.3
Conf. Intv.	3.8	4.2
TER Range	28.8< μx <36.5	
WER Range	21.5< μx <29.9	

	Meteor
Avr.	19.1
Std. Dev.	6.6
Conf. Intv.	1.8
Range	17.3< μx <21.0

	Bleu+
Avr.	48.4
Std. Dev.	27.1
Conf. Intv.	7.5
Range	40.9< μx <55.9

Figure 4.3.1.3: Average Scores of Evaluation Rates for Turkish Complex Sentences on Bing

The most remarkable score statistic shows that the narrow range so the most stable similarity evaluation metric is meteor. Meteor can generate closer scores as generally because of its specific formula.



	Prec.	Recall
Avr.	48.9	48.2
Std. Dev.	15.5	15.2
Conf. Intv.	4.3	4.2
Prec. Range	44.6< μx <53.2	
Recall Range	44.0< μx <52.4	

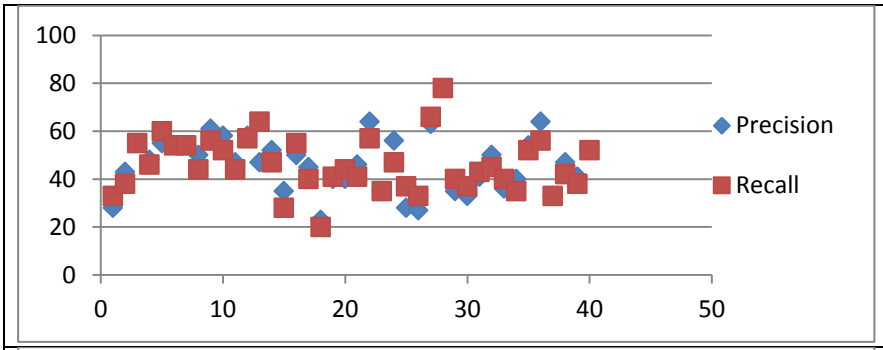
	TER	**WER**
Avr.	33.7	29.2
Std. Dev.	16.6	15.9
Conf. Intv.	4.6	4.4
TER Range	29.1< μx <38.3	
WER Range	24.8< μx <33.6	

	Meteor
Avr.	19.9
Std. Dev.	7.7
Conf. Intv.	2.1
Range	17.8< μx <22.1

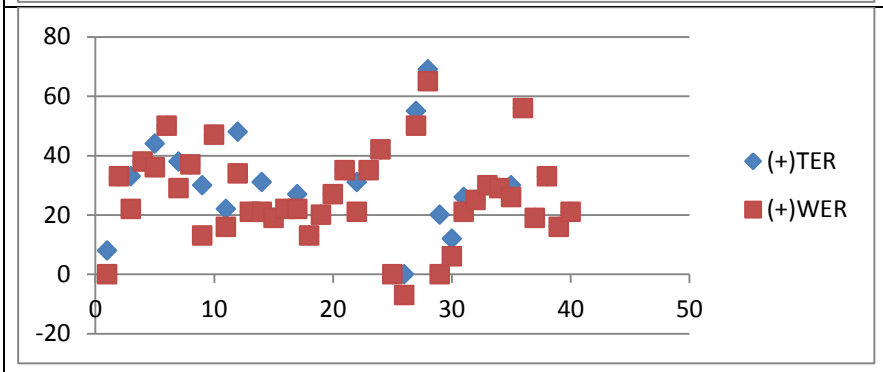
	Bleu+
Avr.	41.8
Std. Dev.	17.9
Conf. Intv.	5.0
Range	36.9< μx <46.8

Figure 4.3.1.4: Average Scores of Evaluation Rates for Turkish Compound Sentences on Bing

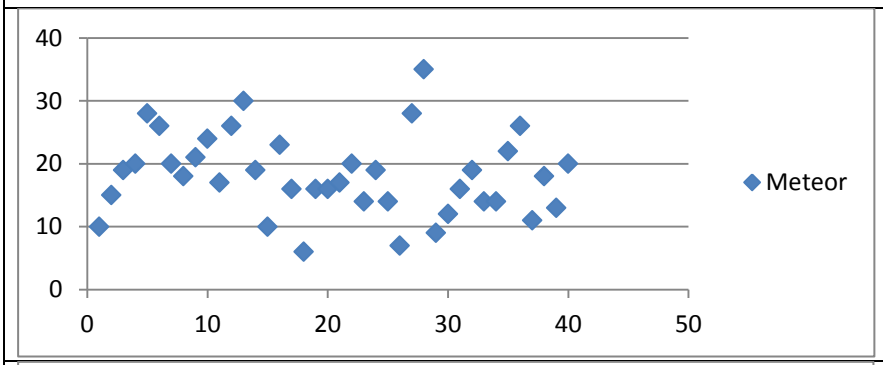
It is seen clearly that the average similarity score of compound sentences translated from English language to Turkish by using with Bing service are in generally ordered ascending from Meteor < WER < TER < Bleu+ < Recall < Precision. And both standard deviation and confidence interval of Bleu+ are at the highest level in other metrics.



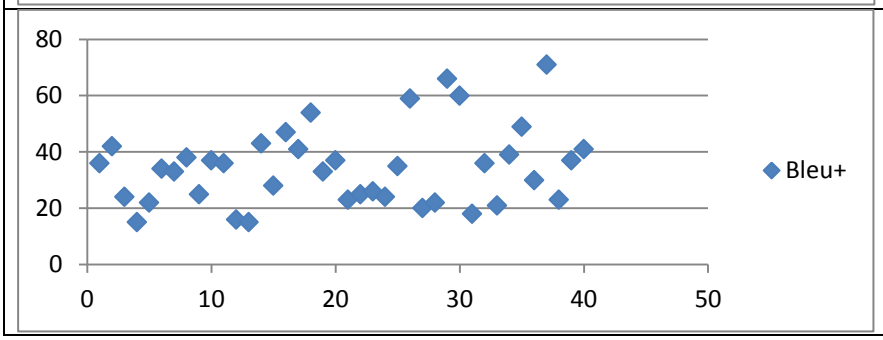
	Prec.	Recall
Avr.	46.7	46.0
Std. Dev.	11.9	11.3
Conf. Intv.	3.7	3.5
Prec. Range	43.0< μx <50.3	
Recall Range	42.5< μx <49.5	



	TER	WER
Avr.	29.8	26.1
Std. Dev.	14.6	15.5
Conf. Intv.	4.5	4.8
TER Range	25.3< μx <34.3	
WER Range	21.3< μx <30.9	



	Meteor
Avr.	18.2
Std. Dev.	6.4
Conf. Intv.	2.0
Range	16.2< μx <20.2



	Bleu+
Avr.	34.5
Std. Dev.	13.8
Conf. Intv.	4.3
Range	30.2< μx <38.8

Figure 4.3.1.5: Average Scores of Evaluation Rates for Turkish Complex-Compound Sentences on Bing

Since specific structure compound-complex sentences and long number of word of that type, Bleu+, TER and WER shows us alignment of word after translation is not better than other sentence structure.

4.3.2 Evaluation Train Test of Bing Service from Turkish to English

The table on the following shows similarity metric rates of translated sources sentences via Bing translation services in terms of given reference sentences in segmentation of sentence structure:

Table 4.3.2.1: Average Scores of Evaluation Rates for All Structures of English Sentences on Bing

#	Sent. Str.	P	R	T	W	M	B	B+
50	**Simple**	50.1	47.2	32.2	25.4	22.5	22.5	42.8
50	**Complex**	54.5	52.7	29.9	15.8	21.1	18.5	30.2
50	**Compound**	56.6	54.3	35.0	27.3	23.5	24.8	32.3
40	**Complex-Compound**	52.0	51.5	31.3	22.4	20.8	19.2	27.2

The Table 4.3.2.1 and Figure 4.3.2.1 show that Bing translation service can translate from Turkish to English better than from Turkish to English language translation. And the structure of compound sentences can be translated better than complex and complex-compound sentences in terms of occurrence and alignment of word according to average evaluation results.

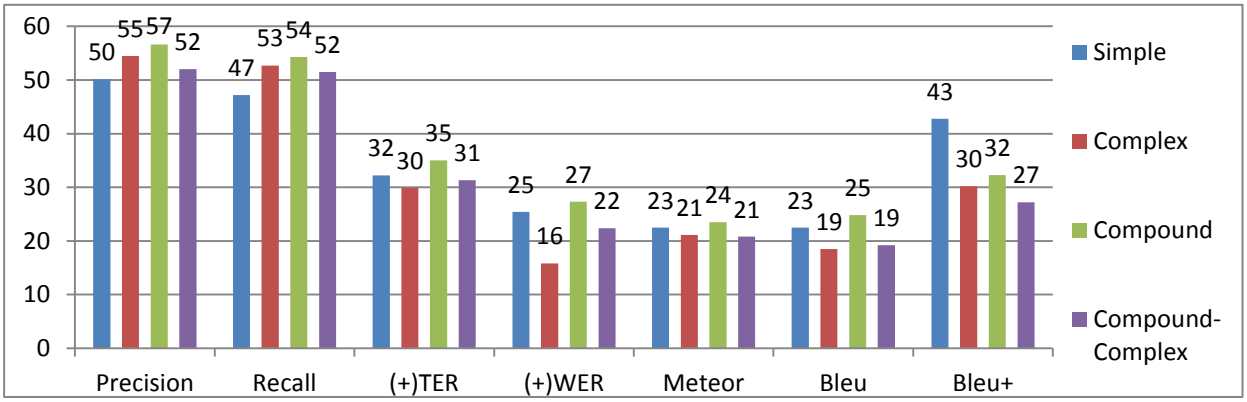
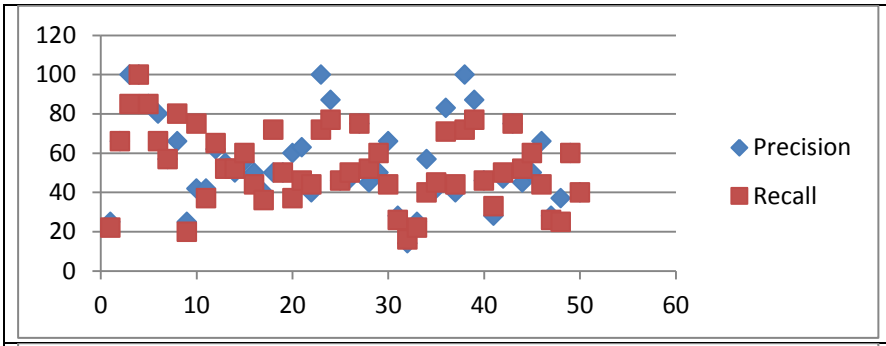
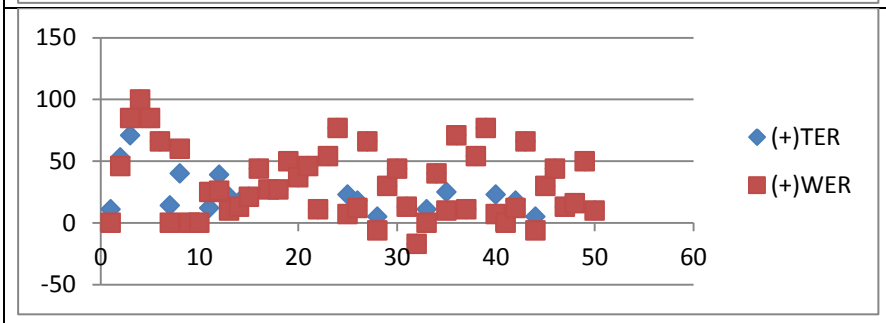


Figure 4.3.2.1: Average Scores of Evaluation Rates for All Structures of English Sentences on Bing

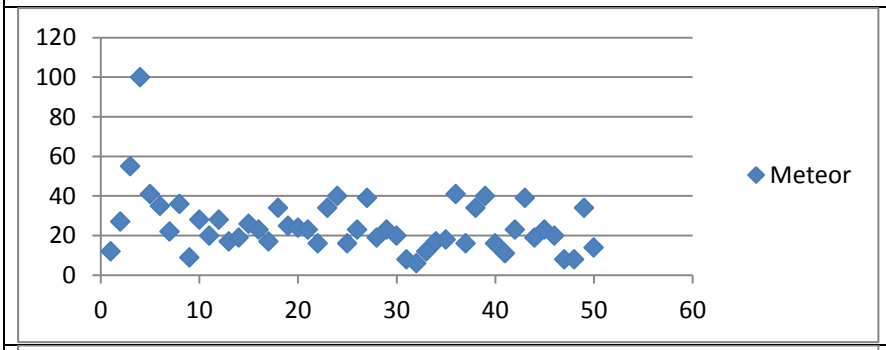
It is seen clearly that Bing translator can translate compound sentences better than the other sentence structure in terms of metrics excluding blue+, but that metric gives evidence about simple sentences with suffix variations.



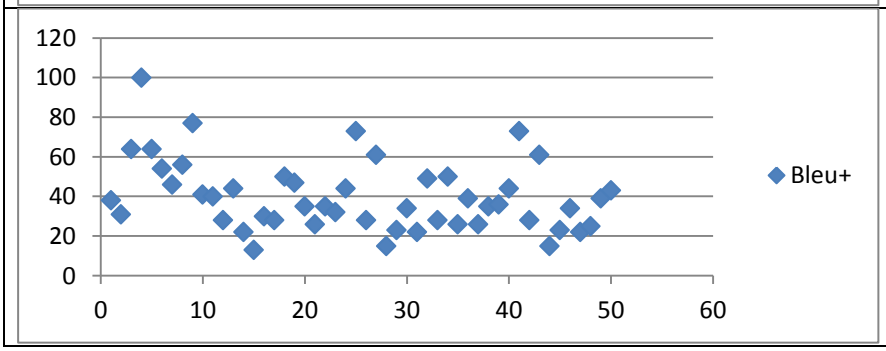
	Precision	Recall
Avr.	50.1	47.2
Std. Dev.	25.1	25.5
Conf. Intv.	7.0	7.1
Prec. Range	43.1< μx <57.0	
Recall range	40.2< μx <54.3	



	TER	WER
Avr.	32.2	25.4
Std. Dev.	29.1	28.9
Conf. Intv.	8.1	8.0
TER Range	24.2< μx <40.3	
WER Range	17.4< μx <33.4	



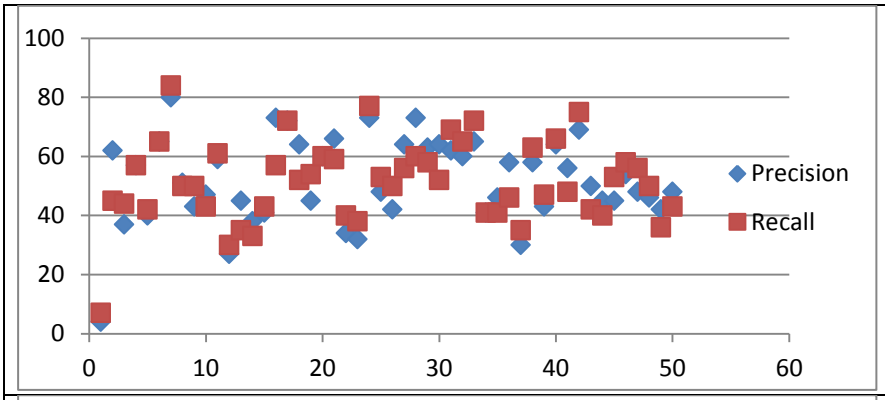
	Meteor
Avr.	22.5
Std. Dev.	21.1
Conf. Intv.	5.9
Range	16.7< μx <28.4



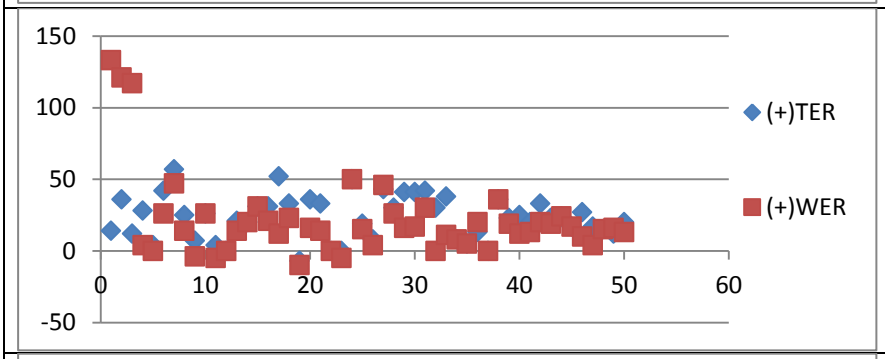
	Bleu+
Avr.	42.8
Std. Dev.	18.4
Conf. Intv.	5.1
Range	37.7< μx <47.9

Figure 4.3.2.2: Average Scores of Evaluation Rates for English Simple Sentences on Bing

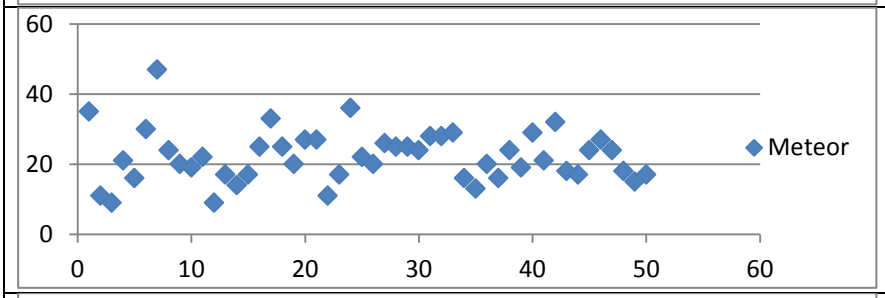
Because of more various synonyms of any word and different type of word group can be used to represent same meaning in English language than Turkish, alignment of word rates are so low.



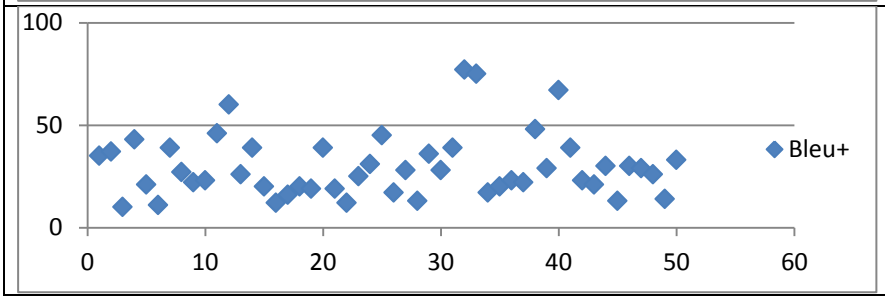
	Precision	Recall
Avr.	54.5	52.7
Std. Dev.	11.1	12.3
Conf. Intv.	3.1	3.4
Prec. Range	51.4< μx <57.6	
Recall Range	49.3< μx <56.2	



	TER	WER
Avr.	29.9	15.8
Std. Dev.	16.4	17.1
Conf. Intv.	4.6	4.7
TER Range	25.4< μx <34.5	
WER Range	11.0< μx <20.5	



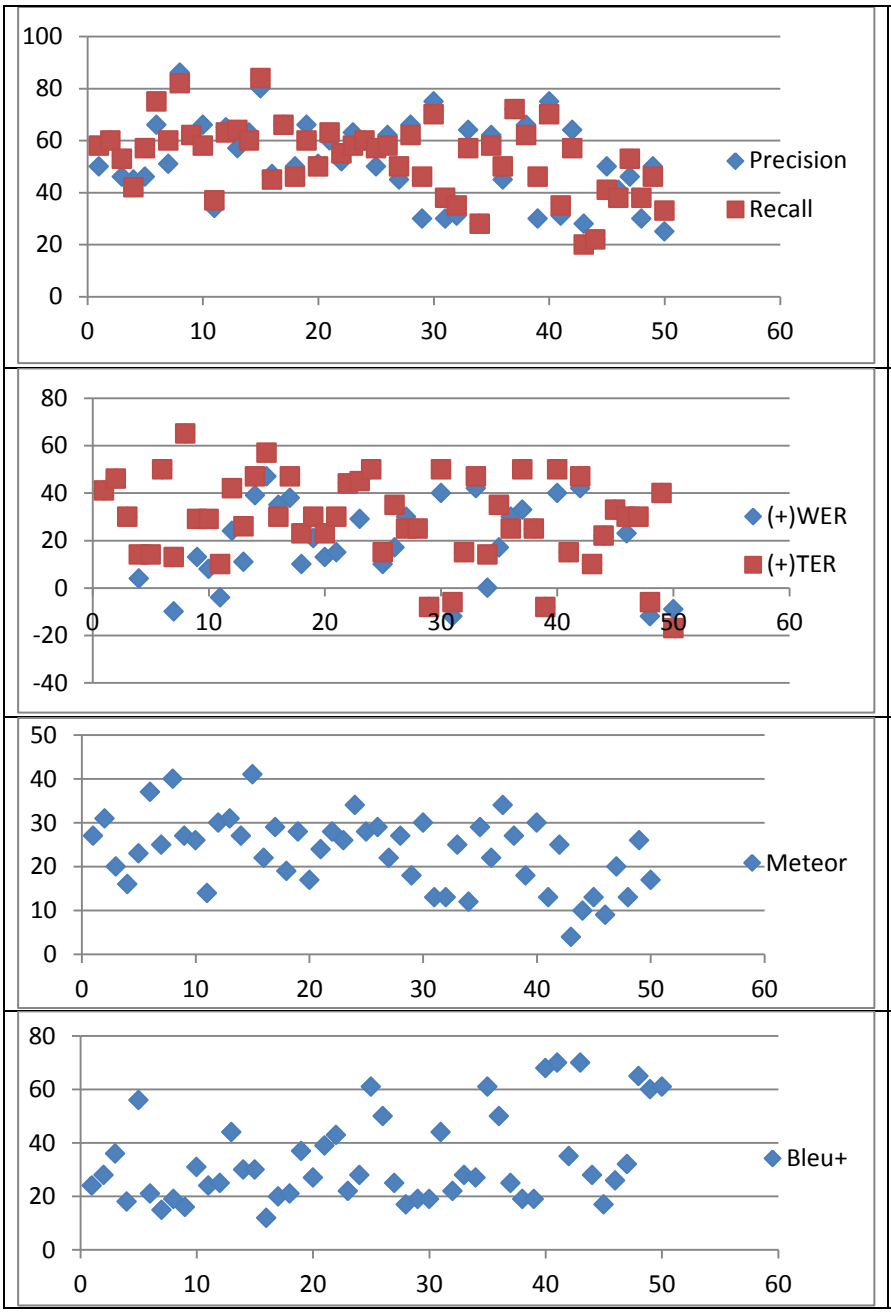
	Meteor
Avr.	21.1
Std. Dev.	7.7
Conf. Intv.	2.1
	18.9< μx <23.2



	Bleu+
Avr.	30.2
Std. Dev.	13.8
Conf. Intv.	3.8
Range	26.3< μx <34.0

Figure 4.3.2.3: Average Scores of Evaluation Rates for English Complex Sentences on Bing

Especially in this evaluation part it is seen that some WER accuracy scores are over 100. So results give evidence about WER accuracy rates failure since insertion or/and deletion process of WER. So TER is preferred since TER is more meaningful instead of WER.



	Prec.	Recall
Avr.	56.6	54.3
Std. Dev.	14.5	12.0
Conf. Intv.	4.0	3.3
Prec. Range	52.6< μx <60.7	
Recall Range	51.0< μx <57.6	

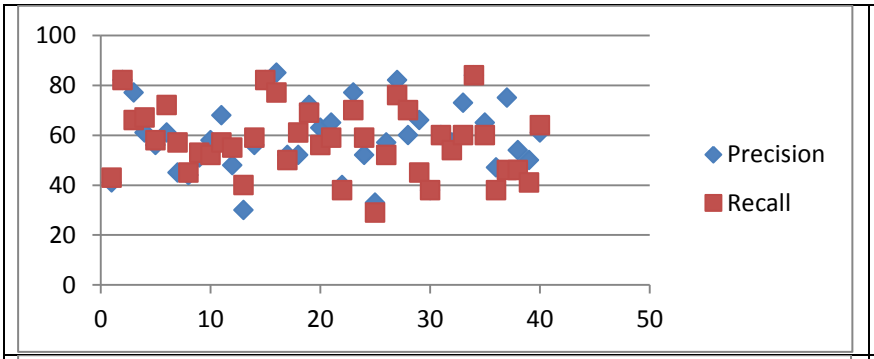
	TER	WER
Avr.	35.0	27.3
Std. Dev.	19.5	20.0
Conf. Intv.	5.4	5.5
TER Range	29.6< μx <40.4	
WER Range	21.8< μx <32.8	

	Meteor
Avr.	23.5
Std. Dev.	7.6
Conf. Intv.	2.1
Range	21.4< μx <25.6

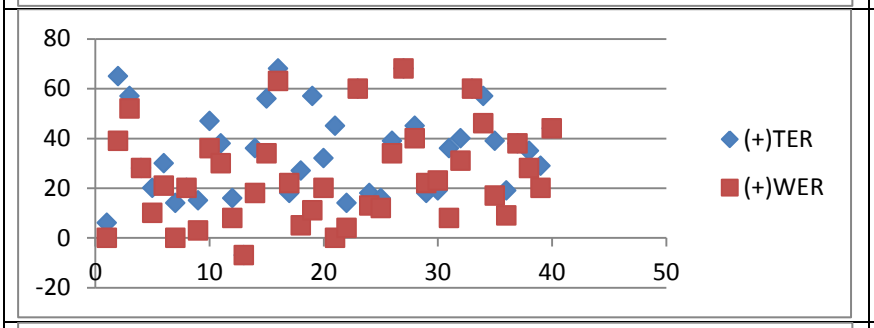
	Bleu+
Avr.	32.3
Std. Dev.	14.2
Conf. Intv.	3.9
Range	28.4< μx <36.3

Figure 4.3.2.4: Average Scores of Evaluation Rates for English Compound Sentences on Bing

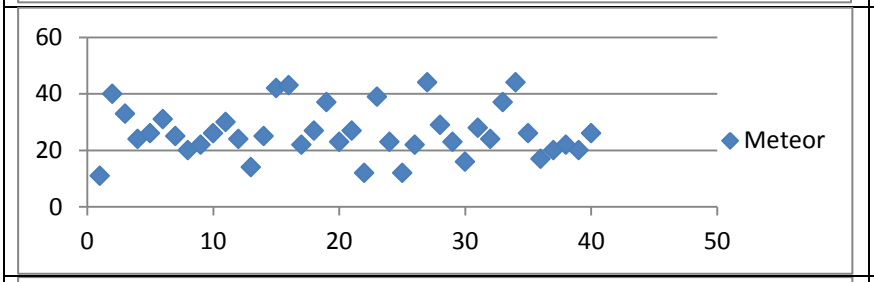
For Bing, compound sentence structure evaluation results show that alignment of occurred word is so mixed. TER and WER accuracy rates are under 0. This proves that both TER and WER formula need some edit and enhancement.



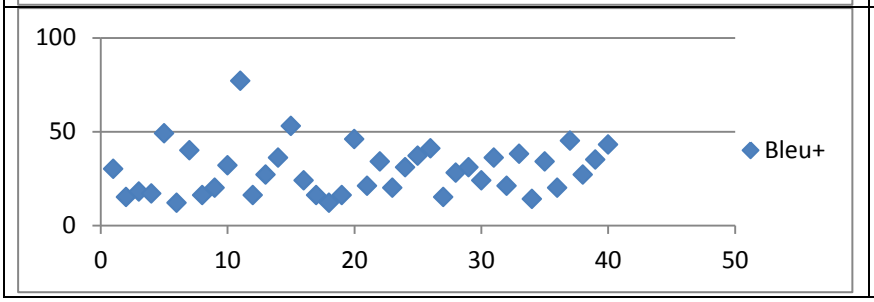
	Prec.	Recall
Avr.	52.0	51.5
Std. Dev.	11.7	10.9
Conf. Intv.	3.6	3.4
Prec. Range	48.4< μx <55.6	
Recall Range	48.1< μx <54.8	



	TER	WER
Avr.	31.3	22.4
Std. Dev.	15.1	16.2
Conf. Intv.	4.7	5.0
TER Range	26.6< μx <36.0	
WER Range	17.4< μx <27.5	



	Mereor
Avr.	20.8
Std. Dev.	6.6
Conf. Intv.	2.0
Range	18.8< μx <22.9



	Bleu+
Avr.	27.2
Std. Dev.	9.6
Conf. Intv.	3.0
Range	24.2< μx <30.1

Figure 4.3.2.5: Average Scores of Evaluation Rates for English Complex-Compound Sentences on Bing

It is seen clearly that since naturally complex and compound sentence are longer and more mixed, alignment of word rates of Bleu, Bleu+, Meteor, etc. are so low in terms of previous structures.

4.4 Automatic Metric Evaluation of Yandex

In this section, Yandex translation service is evaluated by using with language and auto metrics comparatively.

4.4.1 Evaluation Train Test of Yandex Service from English to Turkish

The comparison between Turkish source texts from bilingual corpus and Yandex Turkish translated Turkish texts translated from English source texts by Yandex translation service.

Table 4.4.1.1: Average Scores of Evaluation Rates for All Structures of Turkish Sentences on Yandex

#	Sentence Structure	P	R	T	W	M	B	B+
50	**Simple**	40.2	40.8	22.9	22.4	17.5	14.6	47.7
50	**Complex**	43.9	43.3	25.6	18.1	19.6	14	42
50	**Compound**	41.7	42.6	23.7	22.4	19.3	14	46.1
40	**Complex-Compound**	48.3	46.3	35.4	31.5	22.3	22	36.9

In this evaluation study, Yandex service is evaluated in terms of languages and sentence structures.

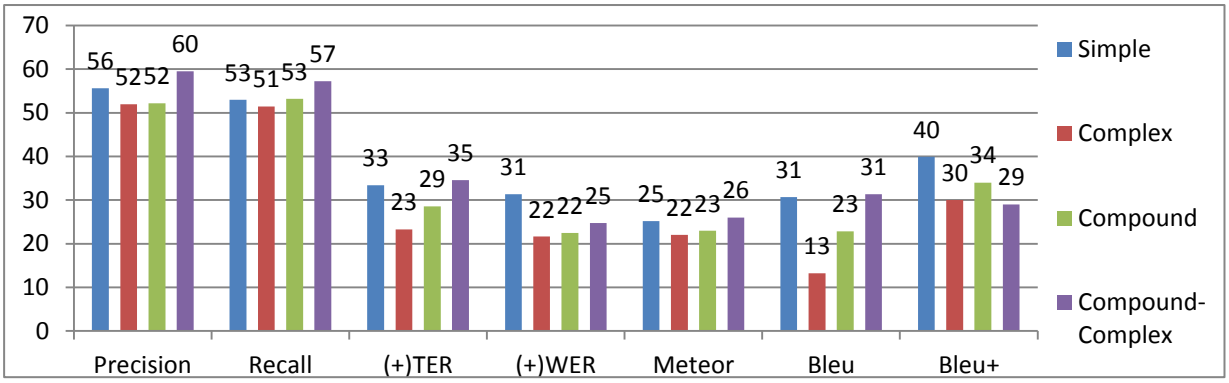
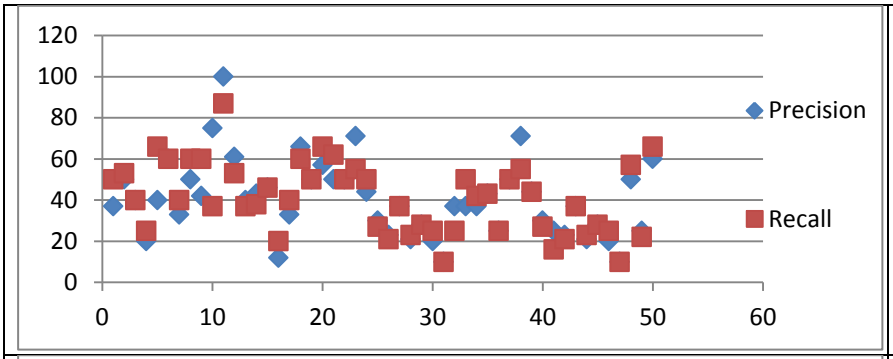
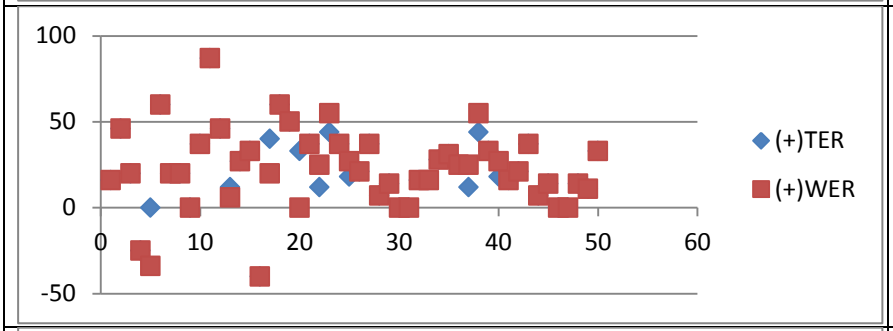


Figure 4.4.1.1: Average Scores of Evaluation Rates for All Structures of Turkish Sentences on Yandex

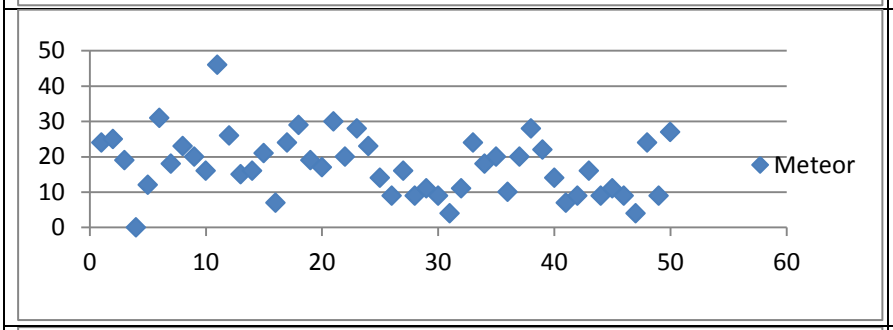
Yandex translation service can translate English text to Turkish language well on compound-complex sentence level word by word. But on root-suffix based evaluation test it is seen clearly by bleu+ metric that Yandex can translate simple and compound sentence much better than other structures. Bleu-Bleu+ comparison show suffix effect. So from English to Turkish translation by Yandex can be almost perform as well as a human approach at suffix level on simple and compound-complex structures. But Bleu+ metric shows compound-complex lower than bleu metric. It may be an application bug.



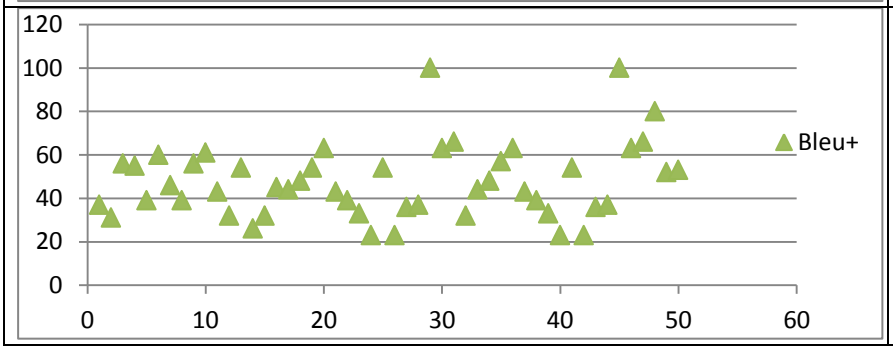
	Prec.	Recall
Avr.	40.2	40.8
Std. Dev.	18.1	17.0
Conf. Intv.	5.0	4.7
Prec. Range	35.2< μx <45.3	
Recall Range	36.1< μx <45.6	



	TER	WER
Avr.	22.9	22.4
Std. Dev.	21.3	23.1
Conf. Intv.	5.9	6.4
TER Range	17.0< μx <28.8	
WER Range	16.0< μx <28.8	



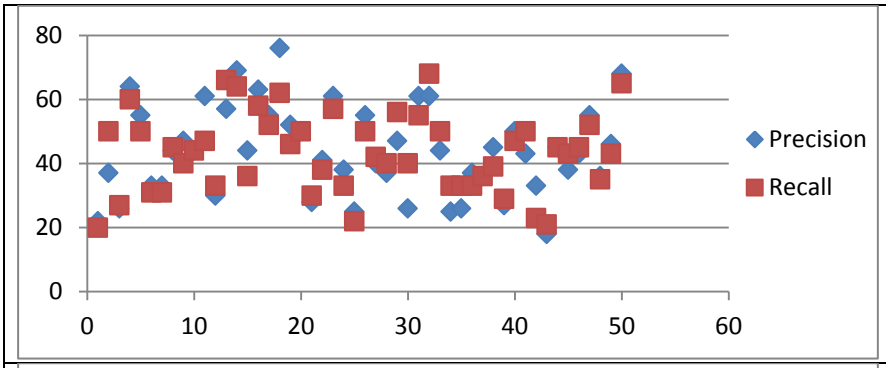
	Meteor
Avr.	17.5
Std. Dev.	8.6
Conf. Intv.	2.4
Range	15.1< μx <19.9



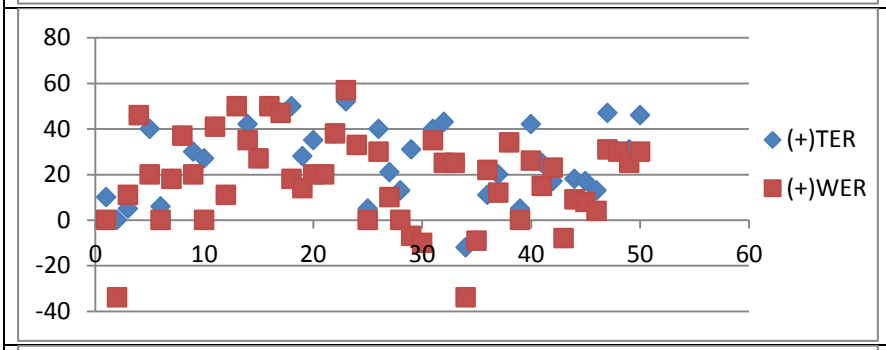
	Bleu+
Avr.	47.7
Std. Dev.	17.1
Conf. Intv.	4.7
Range	42.9< μx <52.4

Figure 4.4.1.2: Average Scores of Evaluation Rates for Turkish Simple Sentences on Yandex

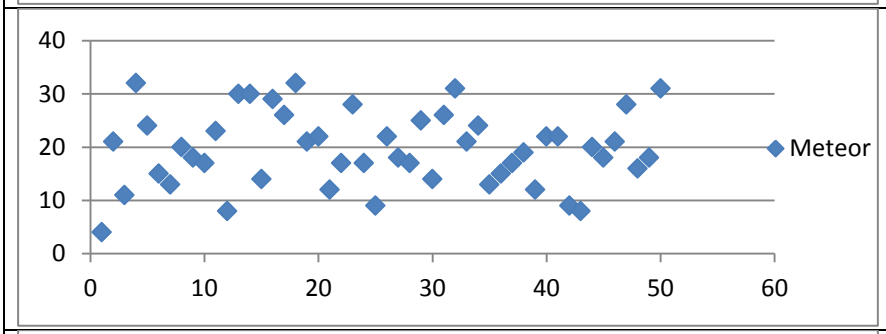
Yandex fascinatingly shows a good performance in translation from English to Turkish language over simple sentence. Although Precision, Recall and other metric average rates on low level, Bleu+ rates are high level since Bleu+ metric calculate root and suffix similarity together after individually measurement.



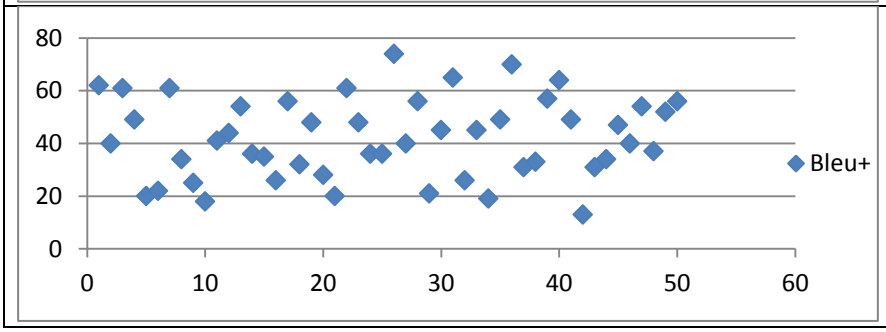
	Prec.	Recall
Avr.	43.9	43.3
Std. Dev.	13.8	12.3
Conf. Intv.	3.82	3.42
Prec. Range	40.1< μx <47.8	
Recall Range	39.9< μx <46.7	



	TER	WER
Avr.	25.6	18.1
Std. Dev.	17.6	19.8
Conf. Intv.	4.88	5.5
TER Range	20.7< μx <30.4	
WER Range	12.6< μx <23.6	



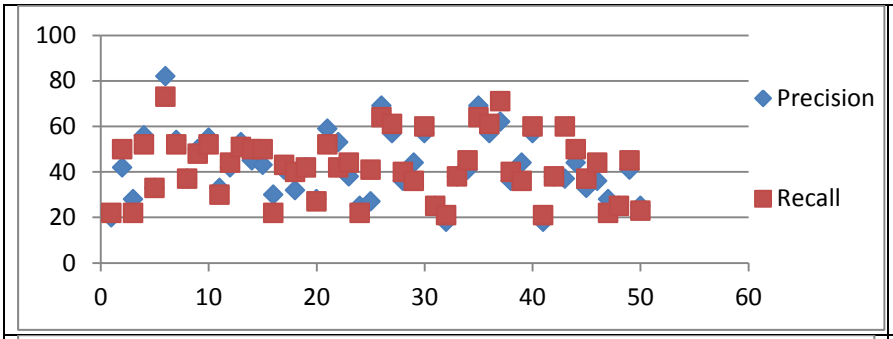
	Meteor
Avr.	19.6
Std. Dev.	6.97
Conf. Intv.	1.93
Range	17.7< μx <21.5



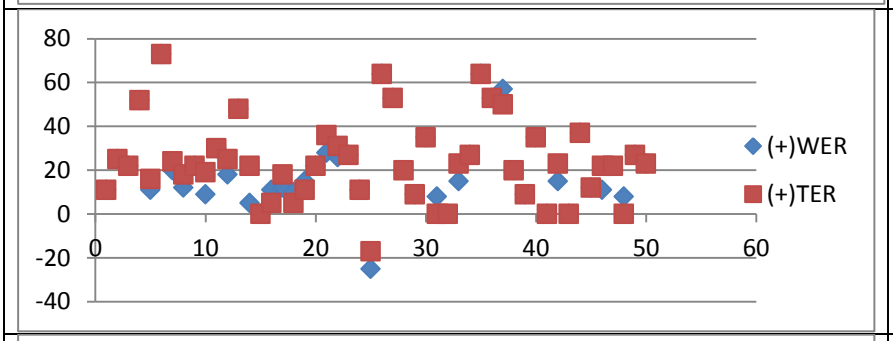
	Bleu+
Avr.	42
Std. Dev.	15.2
Conf. Intv.	4.2
Range	37.8< μx <46.2

Figure 4.4.1.3: Average Scores of Evaluation Rates for Turkish Complex Sentences on Yandex

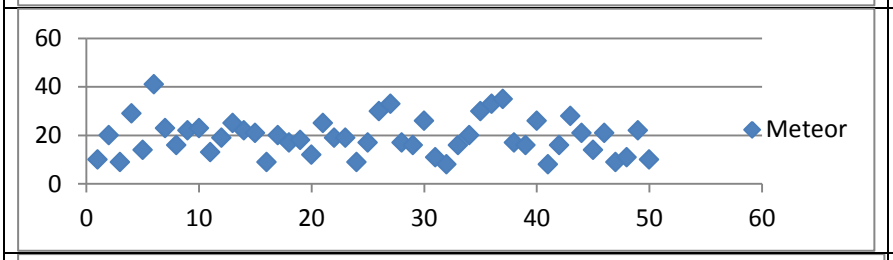
Yandex shows good performance over complex Turkish sentence too. It is unexpected up to now. But evidence shows that Yandex have good performance on translation from English to Turkish language in terms of corpus used.



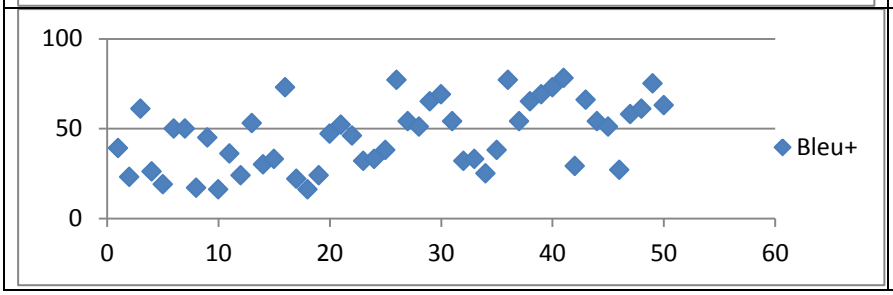
	Prec.	Recall
Avr.	41.7	42.6
Std. Dev.	14.1	14.1
Conf. Intv.	3.92	3.9
Prec. Range	37.7< μx <45.6	
Recall Range	38.7< μx <46.5	



	TER	WER
Avr.	23.7	22.4
Std. Dev.	18.7	19.3
Conf. Intv.	5.19	5.34
TER Range	18.5< μx <28.9	
WER Range	17.0< μx <27.7	



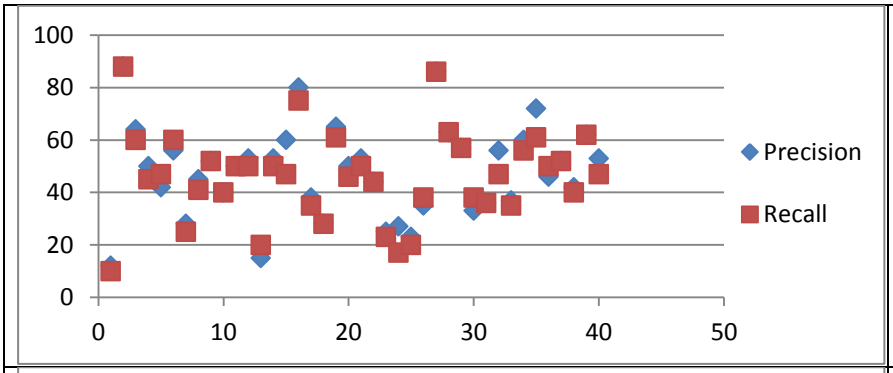
	Meteor
Avr.	19.3
Std. Dev.	7.72
Conf. Intv.	2.14
Range	17.2< μx <21.5



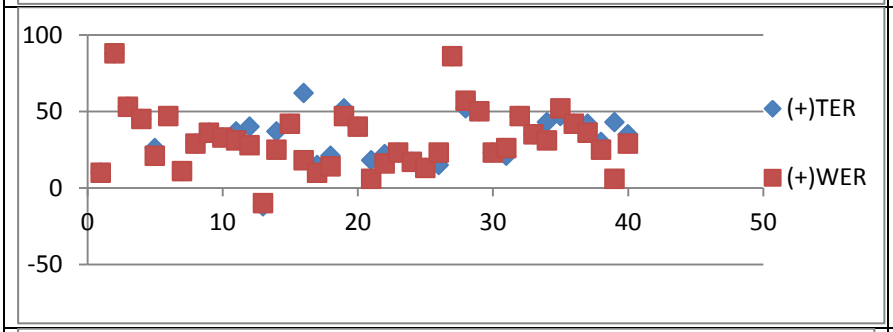
	Bleu+
Avr.	46.1
Std. Dev.	18.7
Conf. Intv.	5.18
Range	40.9< μx <51.2

Figure 4.4.1.4: Average Scores of Evaluation Rates for Turkish Compound Sentences on Yandex

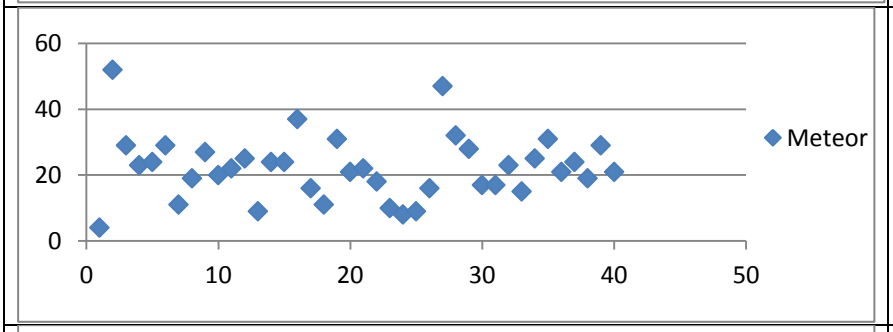
Over Turkish Compound Sentence Yandex service can translate better again. Since compound sentences are consisting of at least two simple sentences, 100% similarity cannot be possible.



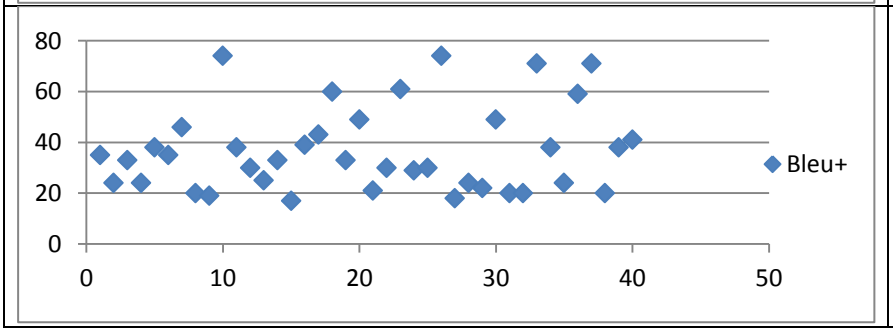
	Prec.	Recall
Avr.	48.3	46.3
Std. Dev.	17.5	17
Conf. Intv.	5.43	5.28
Prec. Range	42.8< μx <53.7	
Recall Range	41.0< μx <51.6	



	TER	WER
Avr.	35.4	31.5
Std. Dev.	19.2	19.9
Conf. Intv.	5.96	6.16
TER Range	29.4< μx <41.4	
WER Range	25.4< μx <37.7	



	Meteor
Avr.	22.3
Std. Dev.	9.75
Conf. Intv.	3.02
Range	19.2< μx <25.3



	Bleu+
Avr.	36.9
Std. Dev.	16.7
Conf. Intv.	5.18
Range	31.7< μx <42.1

Figure 4.4.1.5: Average Scores of Evaluation Rates for Turkish Complex-Compound Sentences on Yandex

As in every test meteor metric shows more densely results than other. On the one hand it is good for stability, and in other hand it cannot show the specifications. So it is worrying about assessment of sentence.

4.4.2 Evaluation Train Test of Yandex Service from Turkish to English

In this section Yandex service was evaluated translation from Turkish to English language over sentence structures individually again. The Table 4.4.2.1 on the following average similarity rates of Yandex service evaluation in terms of sentence structure.

Table 4.4.2.1: Average Scores of Evaluation Rates for All Structures of English Sentences on Yandex

#	Sent. Str.	Precision	Recall	TER	WER	Meteor	Bleu	Bleu+
50	**Simple**	55.6	53.0	33.4	31.3	25.2	30.7	39.9
50	**Complex**	51.98	51.46	23.3	21.62	22	13.2	30
50	**Compound**	52.2	53.2	28.56	22.46	23	22.8	34
40	**Complex-Compound**	59.5	57.25	34.55	24.75	26	31.3	29

The Table 4.4.2.1 gives some evidence that in terms of incising of word alignment design from simple to complex-compound, translation from Turkish to English language Yandex service produces clear results. Success at word matching level with suffixes is very high, but unfortunately meaningful fluent alignment of word and word groups is so low level and it is gradually decreasing from simple to complex compound without only compound sentence.

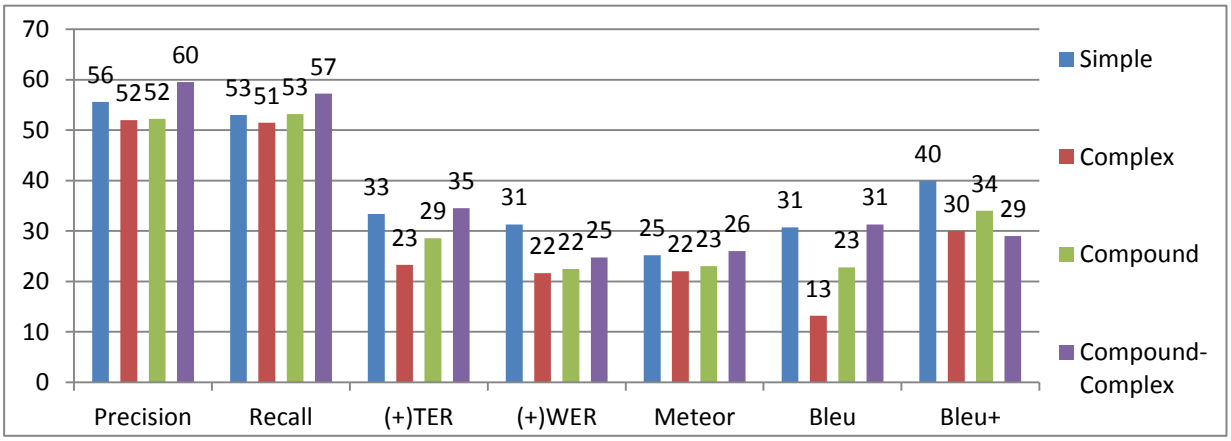
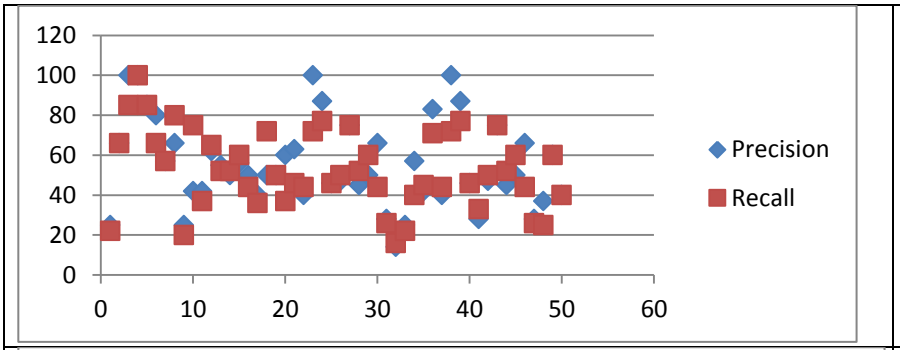
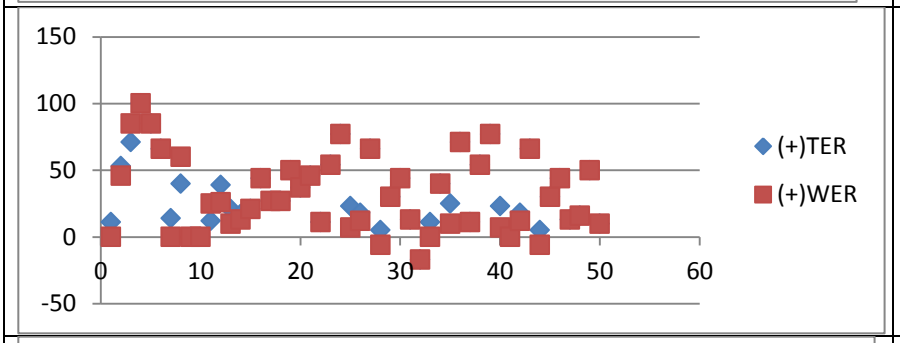


Figure 4.4.2.1: Average Scores of Evaluation Rates for All Structures of English Sentences on Yandex

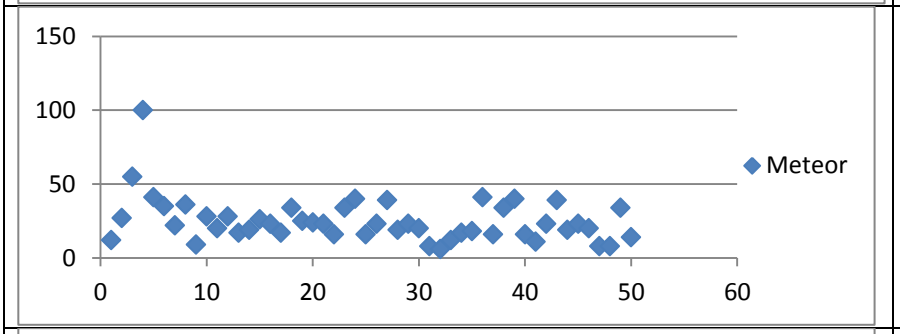
Over complex-compound sentence surprisingly Yandex perform better translation. But for simple and complex sentences, it is seen obviously by bleu+ metric that there are some words in sentence with same word root but different suffixes. Bleu+ earn words to get high correlation with an expert approach about translation if different suffixes are acceptable by the determined rate by the same root or parallel meaning of word root by synonym or phrase.



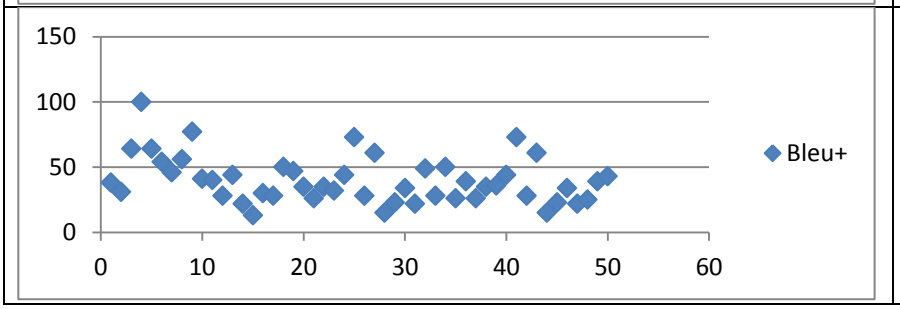
	Prec.	Recall
Avr.	55.6	53.0
Std. Dev.	21.6	19.4
Conf. Intv.	6.0	5.4
Prec. Range	49.6< μx <61.5	
Recall Range	47.6< μx <58.4	



	TER	WER
Avr.	33.4	31.3
Std. Dev.	25.9	28.6
Conf. Intv.	7.2	7.9
TER Range	26.2< μx <40.6	
WER Range	23.4< μx <39.2	



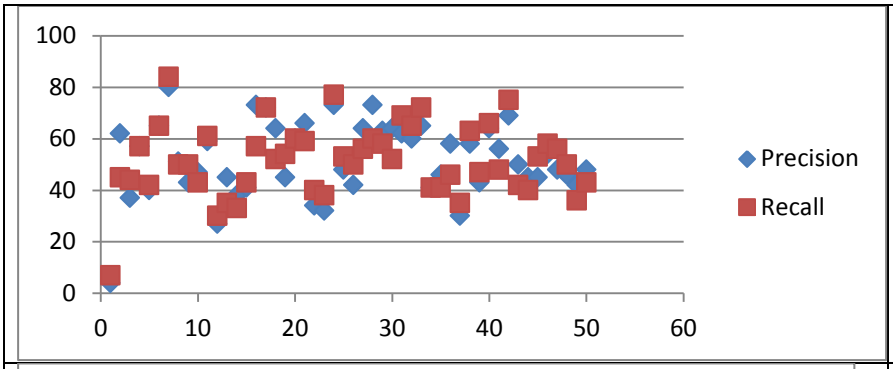
	Meteor
Avr.	25.2
Std. Dev.	15.2
Conf. Intv.	4.2
Range	20.9< μx <29.4



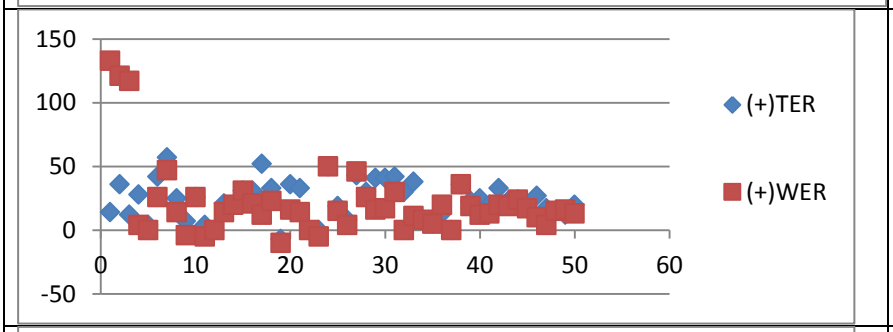
	Bleu+
Avr.	39.9
Std. Dev.	17.8
Conf. Intv.	4.9
Range	35.0< μx <44.9

Figure 4.4.2.2: Average Scores of Evaluation Rates for English Simple Sentences on Yandex

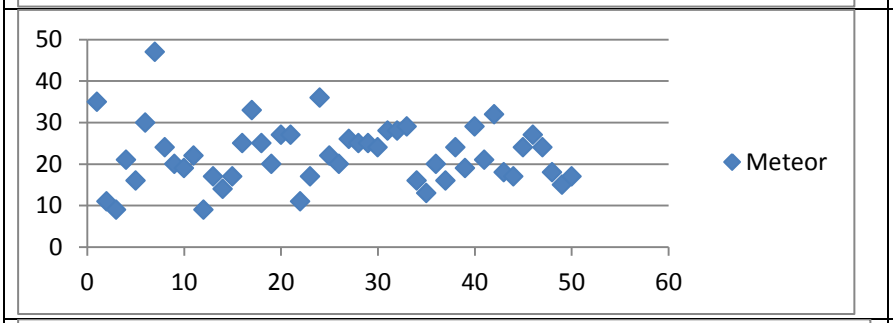
From Turkish to English translation on simple sentence, Yandex gives generally expected rates. Word alignment quality level can be seen obviously by Precision-TER comparison.



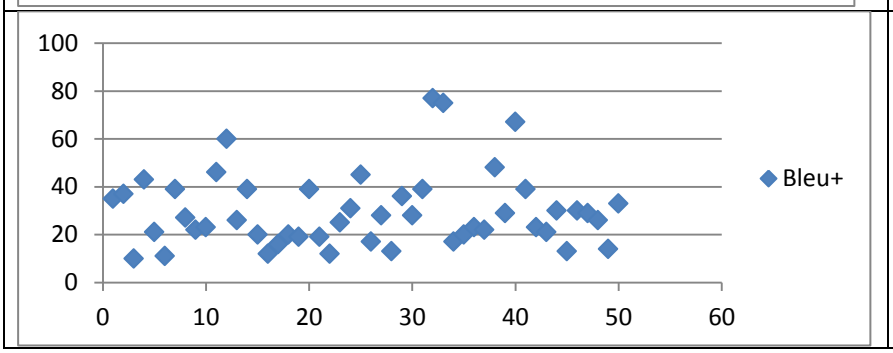
	Prec.	Recall
Avr.	51.9	51.4
Std. Dev.	14.4	13.8
Conf. Intv.	3.9	3.8
Prec. Range	48.0< μx <56.0	
Recall Range	47.6< μx <55.3	



	TER	WER
Avr.	23.3	21.62
Std. Dev.	15.1	29.1
Conf. Intv.	4.1	8.0
TER Range	19.1< μx <27.5	
WER Range	13.5< μx <29.7	



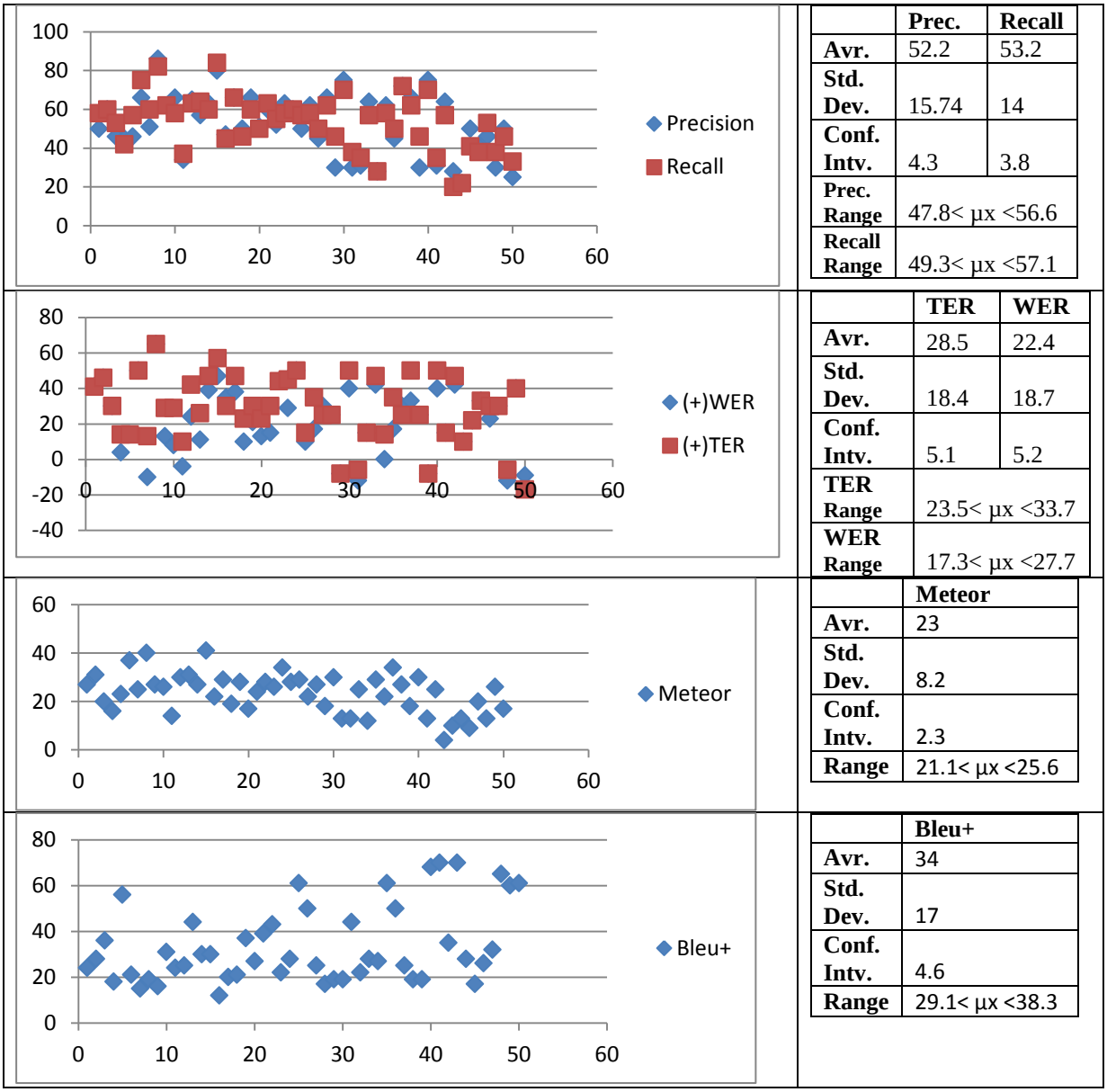
	Meteor
Avr.	22
Std. Dev.	7.4
Conf. Intv.	2
Range	20.1< μx <24.2



	Bleu+
Avr.	30
Std. Dev.	16
Conf. Intv.	4.3
Range	25.6< μx <34.2

Figure 4.4.2.3: Average Scores of Evaluation Rates for English Complex Sentences on Yandex

At Yandex service evaluation is done by complex sentences translated from Turkish to English in this train test. Results can give some interesting idea about complex sentence translated by Yandex. And it's obviously seen that were accuracy rate is giving wrong results. So its approach should be reviewed.



	Prec.	Recall
Avr.	52.2	53.2
Std. Dev.	15.74	14
Conf. Intv.	4.3	3.8
Prec. Range	47.8< μx <56.6	
Recall Range	49.3< μx <57.1	

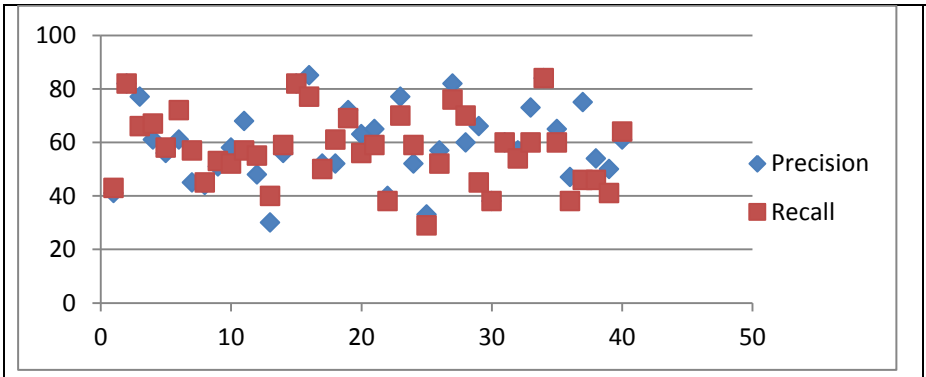
	TER	WER
Avr.	28.5	22.4
Std. Dev.	18.4	18.7
Conf. Intv.	5.1	5.2
TER Range	23.5< μx <33.7	
WER Range	17.3< μx <27.7	

	Meteor
Avr.	23
Std. Dev.	8.2
Conf. Intv.	2.3
Range	21.1< μx <25.6

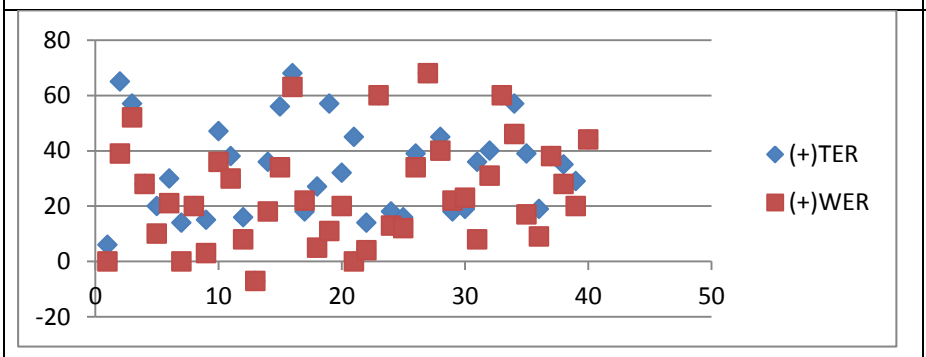
	Bleu+
Avr.	34
Std. Dev.	17
Conf. Intv.	4.6
Range	29.1< μx <38.3

Figure 4.4.2.4: Average Scores of Evaluation Rates for English Compound Sentences on Yandex

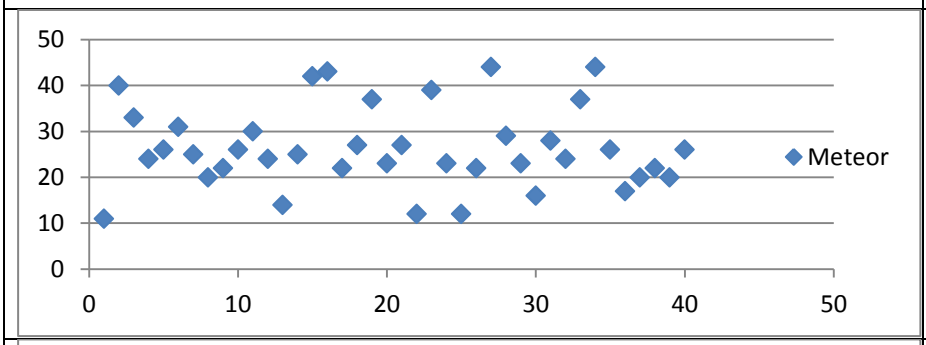
It can be understood that Yandex service can translate compound sentence well at word root, suffix and word alignment.



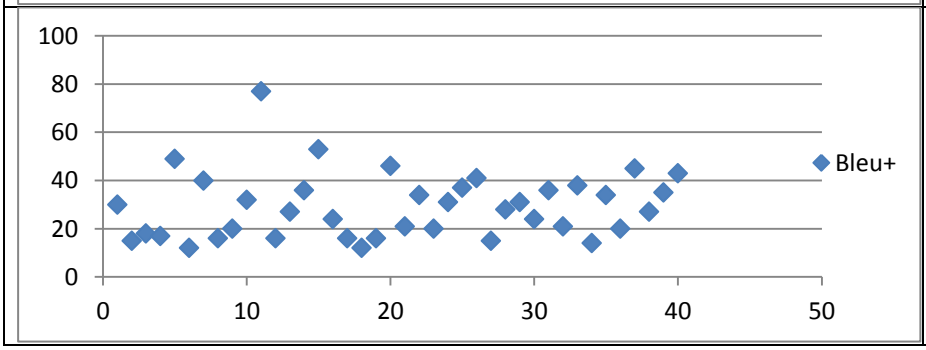
	Prec.	Recall
Avr.	59.5	57.2
Std. Dev.	14.3	13.3
Conf. Intv.	4.4	4.1
Prec. Range	55.0< μx <64.0	
Recall Range	53.1< μx <61.4	



	TER	WER
Avr.	34.5	24.7
Std. Dev.	18.49	19.11
Conf. Intv.	5.7	5.9
TER. Range	28.8< μx <40.3	
WER Range	18.8< μx <30.7	



	Meteor
Avr.	26
Std. Dev.	8.8
Conf. Intv.	2.7
Range	23.7< μx <29.1



	Bleu+
Avr.	29
Std. Dev.	14
Conf. Intv.	4.2
Range	25.0< μx <33.4

Figure 4.4.2.5: Average Scores of Evaluation Rates for English Complex-Compound Sentences on Yandex

Over complex compound Yandex automatic translation service exhibits bad results than previous sentence structure.

4.5 Automatic Verification Test

These verification tests are made to see whether to compare correlation between train tests confidence intervals and verification tests which are done by unsupervised machine learning techniques. On some figures, Precision and WER metrics did not used because they are both similar metrics to recall and TER and recall and TER metric give more meaningful rates. Simple, Complex, Compound and Complex compound Sentence Verify Evaluation Test results with Confidence interval with 95% according to statistical formula are shown on the following:

4.5.1 Verify Evaluation Test for Google English to Turkish

These verification tests aim to estimate how train test results are trustable. Firstly Google train test confidence interval is presented on the below:

Table 4.5.1.1: Confidence Interval Rates of Training Set Test for Turkish Corpus on Google

#	Sent. Str.	Recall	TER	Meteor	Bleu	Bleu+
15	**Simple**	33.5< μx <42.4	7.1< μx <24.3	13.5< μx <18.5	11.6< μx <16.9	44.1< μx <51.8
15	**Complex**	39.0< μx <47.1	20.5< μx <32.0	17.4< μx <22.0	13.0< μx <19.3	32.3< μx <40.8
15	**Compound**	48.9< μx <57.0	29.6< μx <37.4	20.5< μx <24.9	12.4< μx <17.6	34.0< μx <43.8
10	**Complex-Compound**	44.6< μx <52.8	25.4< μx <35.7	17.9< μx <22.6	10.0< μx <17.4	31.6< μx <42.6

Table 4.5.1.1 is giving us confidence interval already trained by using bilingual corpus with Google from Turkish to English translation library. Now the overlap measurements of verification test results are given on the following:

Table 4.5.1.2: Overlap Rates of Verification and Train Test for Turkish Corpus on Google

	Precision	Recall	TER	WER	Meteor	Bleu	Bleu+
Simple	33.3	27	40	40	13	20	6.7
Complex	6.7	13	26.7	26.7	27	20	0
Compound	26.7	27	6.7	6.7	0	20	6.7
Complex-Compound	50.0	30	10	20	40	30	40

Precision is giving so closer rates to recall and WER is similar to TER and TER metric is more reliable than WER. So I reduced metric array as recall, TER, Meteor, Bleu and Bleu+.

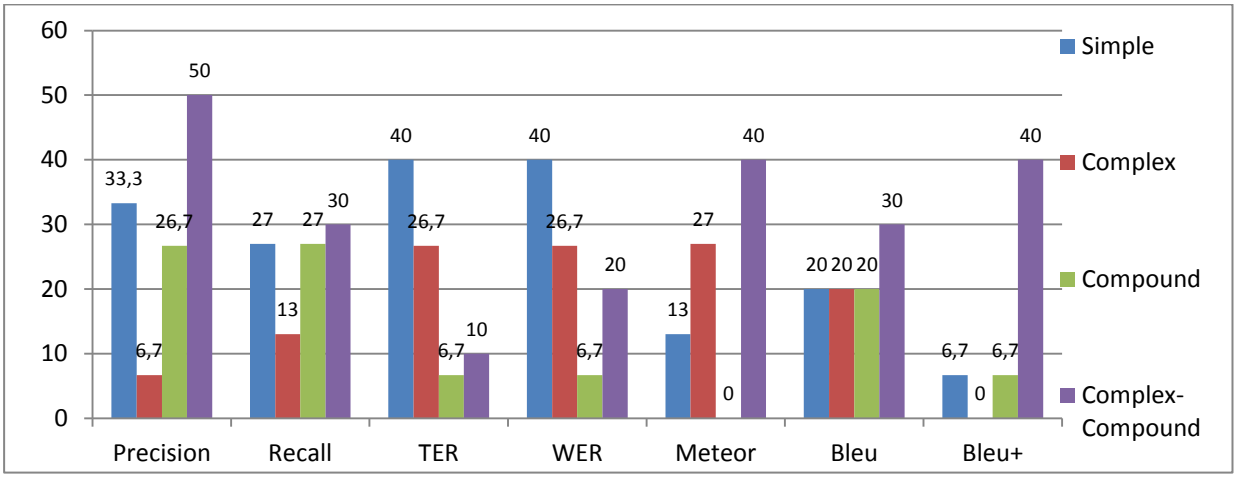


Figure 4.5.1.1: Overlap Rates of Verification and Train Test for Turkish Corpus on Google

According to Figure 4.5.1.1 Meteor and Bleu+ rates give more stable similarity rates over complex-compound sentences. In addition, TER metric is suitable for simple sentence. As seen on Figure 4.5.1.1, for instance, almost all one of two complex-compound sentences may be in 50% precision rates.

4.5.2 Verify Evaluation Test for Google Turkish to English

Table 4.5.2.1 contains confidence intervals with 95% percent as considered in section 3.5.

Table 4.5.2.1: Confidence Interval Rates of Training Set Test for English Corpus on Google

#	Sent. Str.	Recall	TER	Meteor	Bleu	Bleu+
15	**Simple**	51.7< μx <62.7	35.5< μx <50.3	23.9< μx <33.9	25.3< μx <38.5	40.7< μx <50.3
15	**Complex**	52.8< μx <60.7	29.9< μx <38.3	22.4< μx <27.3	21.4< μx <28.2	25.4< μx <33.5
15	**Compound**	52.2< μx <59.9	33.8< μx <42.3	22.9< μx <27.7	21.9< μx <29.4	28.1< μx <34.6
10	**Complex-Compound**	49.0< μx <57.1	30.4< μx <40.4	21.0< μx <26.0	19.3< μx <28.5	24.1< μx <30.9

This train set confidence interval Table 4.5.2.2 is verified in terms of validation set and overlap rate results are represented on the following:

Table 4.5.2.2: Overlap Rates of Verification and Train Test for English Corpus on Google

#	Sent. Str.	Precision	Recall	TER	WER	Meteor	Bleu	Bleu+
15	**Simple**	6.7	27	20	13	33	13	40
15	**Complex**	20	6.7	0	0	6.7	0	0
15	**Compound**	0	0	6.7	0	13	6.7	0
10	**Complex-Compound**	10	0	10	40	10	40	30

Following figure displays the table above:

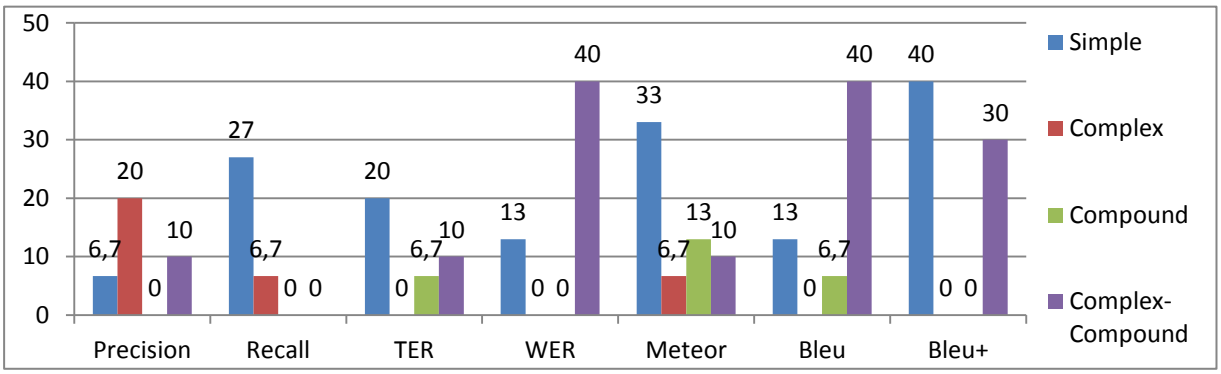


Figure 4.5.2.1: Overlap Rates of Verification and Train Test for English Corpus on Google

In this Figure 4.5.2.1 results shows that there is no capability to estimate a confidence interval over compound sentences generally. With some metric such as TER, WER, Bleu and Bleu+, we cannot validate any estimation about stability of confidence interval.

4.5.3 Verify Evaluation Test for Bing from English to Turkish

Table 4.5.3.1 shows us train set confidence intervals of auto metric evaluation rates between reference and candidate sentence. Candidate sentences are coming from Bing translation service translated from English source texts.

Table 4.5.3.1: Confidence Intervals Rates of Training Set Test for Turkish Corpus on Bing

#	Sent. Str.	Precision	Recall	TER	WER	Meteor	Bleu	Bleu+
15	Simple	37.2< μx <49.4	37.8< μx <49.4	19.1< μx <35.1	19.2< μx <34.7	15.7< μx <24.0	12.4< μx <22.1	57.0< μx <72.1
15	Complex	46.9< μx <53.7	44.6< μx <50.9	28.8< μx <36.5	21.5< μx <29.9	17.3< μx <21.0	8.3< μx <12.3	40.9< μx <55.9
15	Compound	44.6< μx <53.2	44.0< μx <52.4	29.1< μx <38.3	24.8< μx <33.6	17.8< μx <22.1	10.0< μx <16.2	36.9< μx <46.8
10	Complex-Compound	43.0< μx <50.3	42.5< μx <49.5	25.3< μx <34.3	21.3< μx <30.9	16.2< μx <20.2	8.3< μx <12.7	30.2< μx <38.8

The Table 4.5.3.2 shows us validation test results of Bing service train set metric confidence interval over English to Turkish sentences separated by structures.

Table 4.5.3.2: Overlap Rates of Verification and Train Test for Turkish Corpus on Bing

#	Sentence Structure	Precision	Recall	TER	WER	Meteor	Bleu	Bleu+
15	Simple	26.6	0	20	33	40	27	33
15	Complex	6.6	27	6.7	6.7	20	13	13
15	Compound	13.3	27	33	40	13	6.7	20
10	Complex-Compound	30	30	30	50	40	20	10

Moreover, this table is give rates of validation of estimation confidence interval figured on the following:

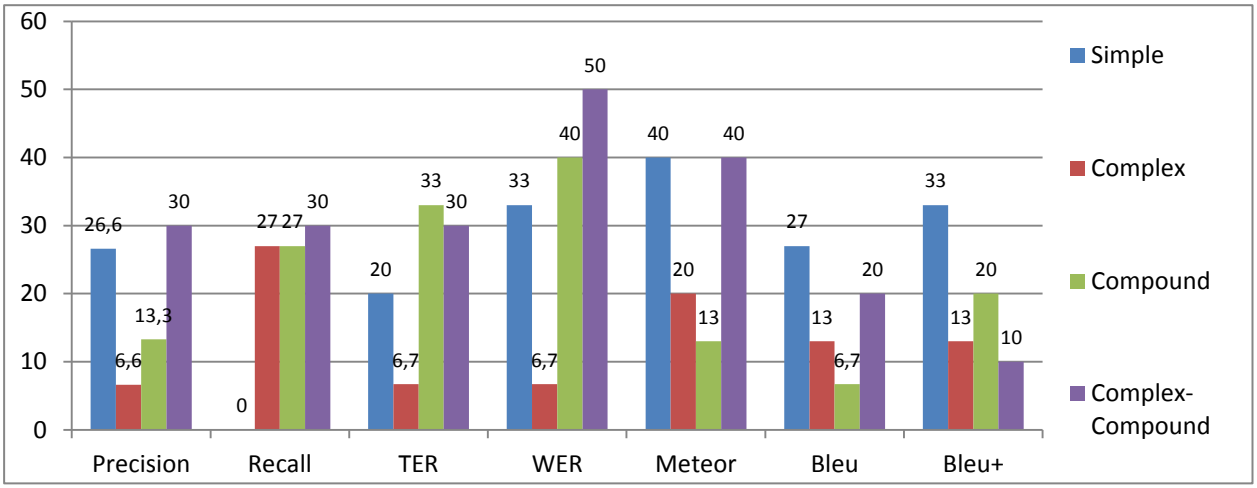


Figure 4.5.3.1: Overlap Rates of Verification and Train Test for Turkish Corpus on Bing

Metrics are representing the validation overlap test rates of training confidence intervals. According to the figure, excluding of recall, simple sentence metric overlap rates are greater than com inside estimation range. In addition, WER metric confidence rate is 50% consistent with validation set rates.

4.5.4 Verify Evaluation Test for Bing Turkish to English

In this section validation results of Bing translation service's evaluation results are exhibited over Turkish to English type approach with 4 main sentence structures.

Table 4.5.4.1: Confidence Intervals Rates of Training Set Test for English Corpus on Bing

#	Sent. Str.	Precision	Recall	TER	WER	Meteor	Bleu	Bleu+
15	Simple	43.1< μx <57.0	40.2< μx <54.3	24.2< μx <40.3	17.4< μx <33.4	16.7< μx <28.4	16.4< μx <28.6	37.7< μx <47.9
15	Complex	51.4< μx <57.6	49.3< μx <56.2	25.4< μx <34.5	11.0< μx <20.5	18.9< μx <23.2	15.3< μx <21.7	26.3< μx <34.0
15	Compound	52.6< μx <60.7	51.0< μx <57.6	29.6< μx <40.4	21.8< μx <32.8	21.4< μx <25.6	21.3< μx <28.3	28.4< μx <36.3
10	Complex-Compound	48.4< μx <55.6	48.1< μx <54.8	26.6< μx <36.0	17.4< μx <27.5	18.8< μx <22.9	15.5< μx <22.9	24.2< μx <30.1

Firstly, I gave similarity confidence interval of sentence translation average values before validation test results. And then verification test over train test confidence interval table on the following:

Table 4.5.4.2: Overlap Rates of Verification and Train Test for English Corpus on Bing

#	Sent. Str.	Precision	Recall	TER	WER	Meteor	Bleu	Bleu+
15	Simple	33	20	47	47	40	0	13
15	Complex	0	6.7	20	6.7	0	13	6.7
15	Compound	0	0	0	0	0	20	6.7
10	Complex-Compound	10	30	20	10	10	10	30

Verification test of train confidence interval rates are figured on the following:

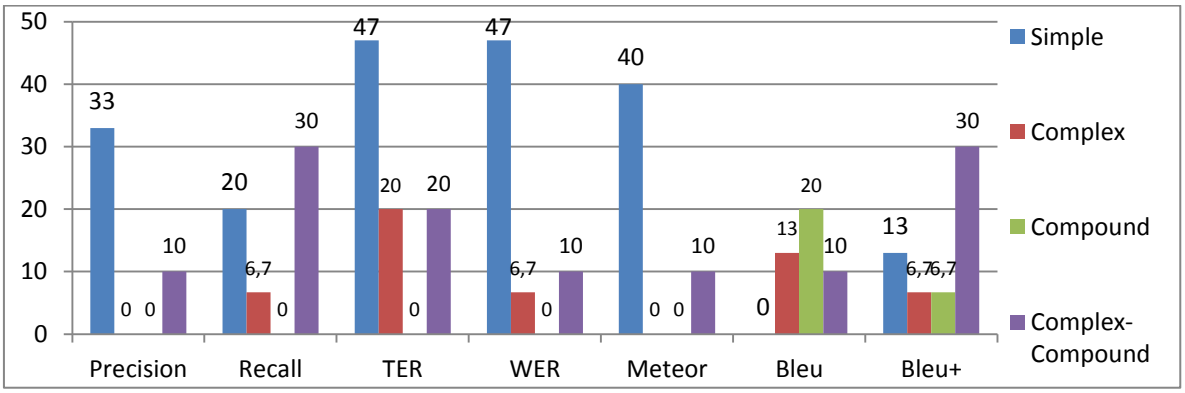


Figure 4.5.4.1: Overlap Rates of Verification and Train Test for English Corpus on Bing

The tables and figures reflect validation test results that give information about metric rates with 95% confidence interval of estimation rates or check to be sure whether also calculated general confidence intervals especially determined unique to sentence structure individually.

4.5.5 Verify Evaluation Test for Yandex from English to Turkish

This table consider about Auto metric similarity evaluation confidence interval between English Reference texts of corpus and candidate sentences which are coming from Yandex translation services translated from English source texts of corpus. Confidence intervals are calculated by statistical standards as shown in section 3.5.

Table 4.5.5.1: Confidence Intervals Rates of Training Set Test for Turkish Corpus on Yandex

#	Sent. Str.	Precision	Recall	TER	WER	Meteor	Bleu	Bleu+
15	Simple	35.2< μx <45.3	36.1< μx <45.6	17.0< μx <28.8	16.0< μx <28.8	15.1< μx <19.9	12.1< μx <17.1	42.9< μx <52.4
15	Complex	40.1< μx <47.8	39.9< μx <46.7	20.7< μx <30.4	12.6< μx <23.6	17.7< μx <21.5	11.9< μx <16.9	37.8< μx <46.2
15	Compound	37.7< μx <45.6	38.7< μx <46.5	18.5< μx <28.9	17.0< μx <27.7	17.2< μx <21.5	11.1< μx <17.8	40.9< μx <51.2
10	Complex-Compound	42.8< μx <53.7	41.0< μx <51.6	29.4< μx <41.4	25.4< μx <37.7	19.2< μx <25.3	16.5< μx <26.7	31.7< μx <42.1

The table shows that similarity evaluation results are been almost in same range for different metric especially Yandex Turkish sentence. The most stable metrics are meteor and bleu on complex sentence structure with 40% overlap rates.

Table 4.5.5.2: Overlap Rates of Verification and Train Test for Turkish Corpus on Yandex

#	**Sent. Str.**	Precision	Recall	TER	WER	Meteor	Bleu	Bleu+
15	**Simple**	20.0	6.7	6.7	0.0	20.0	20.0	6.7
15	**Complex**	13.3	0.0	13.3	33.3	40.0	40.0	13.3
15	**Compound**	6.7	0.0	6.7	13.3	26.7	33.3	33.3
10	**Complex-Compound**	30.0	20.0	20.0	20.0	20.0	10.0	20.0

With the automatic validation test is arguable, its results of English to Turkish evaluation of Yandex service is shown on the table above and figure below:

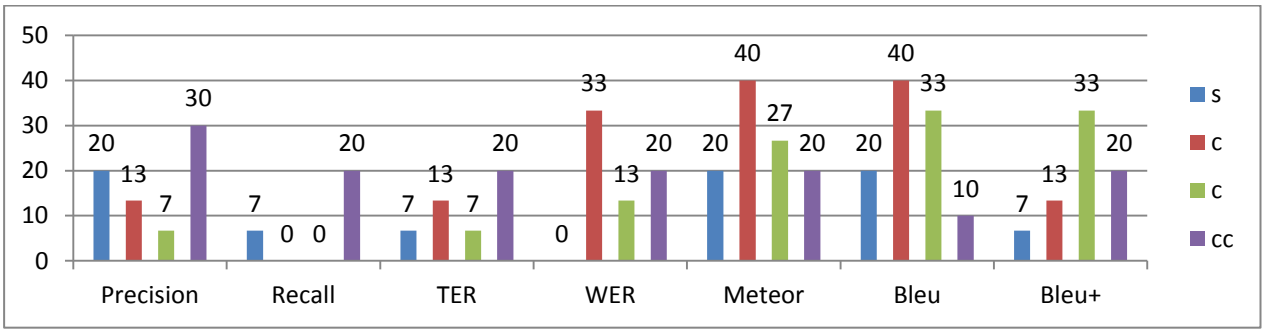


Figure 4.5.5.1: Overlap Rates of Verification and Train Test for Turkish Corpus on Yandex

Figure 4.5.5.1 is to give some facts about Yandex translation service ability that compound sentences can be translated highly in a confidence interval much more than other sentence structure.

4.5.6 Verify Evaluation Test for Yandex from Turkish to English

In this sub section, Turkish to English evaluation of Yandex translations services over sentence structures are tried whether their confidence intervals can be validated or not.

Table 4.5.6.1: Confidence Intervals Rates of Training Set Test for English Corpus on Yandex

#	Sent. Str.	Precision	Recall	TER	WER	Meteor	Bleu	Bleu+
15	**Simple**	49.6< μx <61.5	47.6< μx <58.4	26.2< μx <40.6	23.4< μx <39.2	20.9< μx <29.4	24.8< μx <36.6	35.0< μx <44.9
15	**Complex**	48.0< μx <56.0	47.6< μx <55.3	19.1< μx <27.5	13.5< μx <29.7	20.1< μx <24.2	5.0< μx <21.4	25.6< μx <34.2
15	**Compound**	47.8< μx <56.6	49.3< μx <57.1	23.5< μx <33.7	17.3< μx <27.7	21.1< μx <25.6	19.1< μx <26.6	29.1< μx <38.3
10	**Complex-Compound**	55.0< μx <64.0	53.1< μx <61.4	28.8< μx <40.3	18.8< μx <30.7	23.7< μx <29.1	25.8< μx <36.9	25.0< μx <33.4

The Table 4.2.10 above is representing confidence interval. And then validation test results of Turkish to English evaluation results of Yandex translation service over sentence structure.

Table 4.5.6.2: Overlap Rates of Verification and Train Test for English Corpus on Yandex

#	Sent. Str.	Precision	Recall	TER	WER	Meteor	Bleu	Bleu+
15	**Simple**	53.3	33.3	33.3	33.3	40.0	13.3	26.7
15	**Complex**	26.7	6.7	13.3	20.0	26.7	53.3	0.0
15	**Compound**	0.0	6.7	0.0	0.0	0.0	20.0	13.3
10	**Complex-Compound**	30.0	30.0	60.0	60.0	50.0	0.0	20.0

The Table 4.2.11 above and Figure 4.2.9 are reflecting validation rates of confidence intervals of evaluation Yandex translation service.

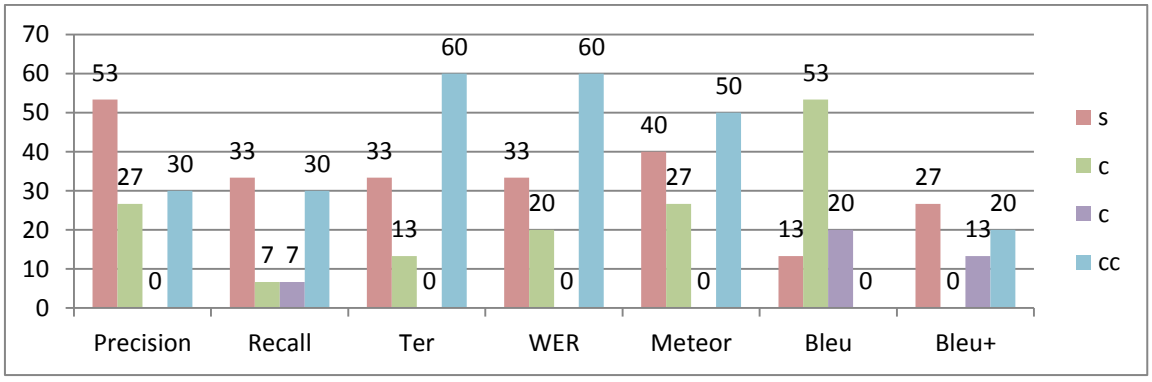


Figure 4.5.6.1: Overlap Rates of Verification and Train Test for English Corpus on Yandex

Figure is evidence to can not to say anything exactly about confidence intervals since subject and word alignment of almost all sentence can be different. So this estimation ranges are not true every time, But they may give useful ideas.

4.6 Human Evaluation (Judgment)

In this section, automatic machine evaluations are verified with human judgments respectively. There are 6 different English teachers judged 32 sentence pairs by using with Costa MT tool.

Instead of matching web translation services, automatic machine translation evaluation methods were compared with human evaluation approaches such as adequacy and fluency to verify automatic evaluation methods' compatibility.

4.6.1 Human Evaluation over Turkish Train Subset

Machine Evaluation versus Human Evaluation is done for validation/verify by human. These evaluation tests results are coming from Asiya evaluation tool and human resource who are English teachers. We have created a small subset of big corpus scaled tables which divided 3 set including Google, Bing and Yandex translation outputs of same source sentence. English to Turkish translation evaluation is made over 32 sentences by matching expert mind on the following:

Table 4.6.1.1: Auto Metric vs. Human Evaluation Comparatively Tests Rates

Set / Methods	Precision	Recall	TER	WER	Meteor	Bleu	Bleu+	Fluency	Adequacy
Google	35	39	10	8	16	14	42	55	59
Bing	58	55	44	41	24	22	51	65	70
Yandex	45	45	25	20	18	15	43	42	45

The Table 4.6.1.1 and its figure, figure 4.6.1.1 shows us similarity evaluation results of random selected sentences comes from Turkish source corpus and translated sentence from English to Turkish by using online translation services, Google, Bing and Yandex. These results gives some evidence about comparison of auto metric and manual, or we can say human made metric, evaluation such as Precision, Recall, Meteor, Bleu, etc. and Fluency and Adequacy.

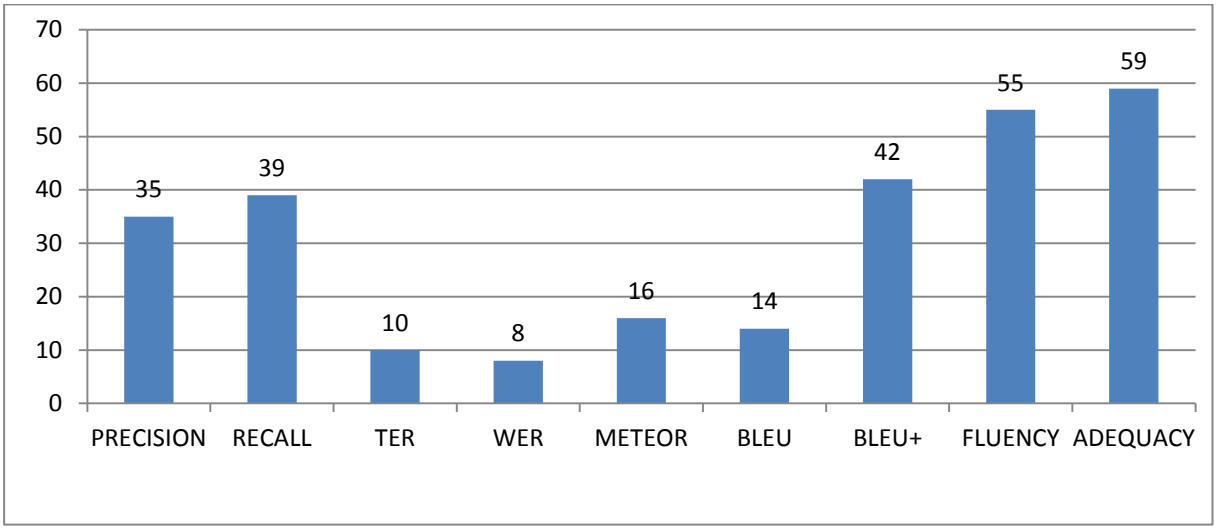


Figure 4.6.1.1: Comparison of both Automatic and Human Evaluation on Google

Figure 4.6.1.1 mainly shows the comparison of machine and human evaluation rates over mixed type of 32 Turkish sentences. When fluency is showing 55 and Adequacy is 59 percent of average similarity rate, Recall is at 39, Precision is at 35 and Meteor and Bleu are on 16 and 14 similarity level. So we can say that Google generally select true words and it aligns them well. But some of suffix determination is missing.

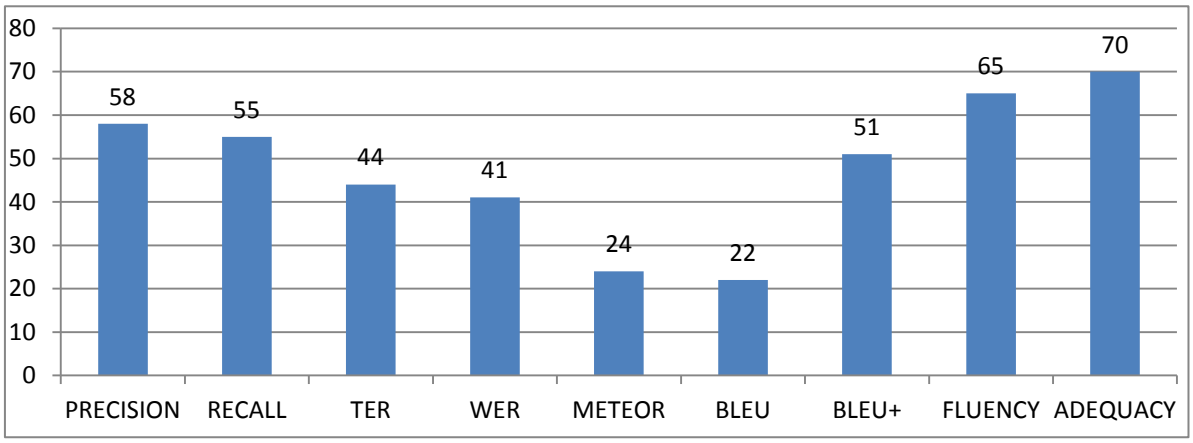


Figure 4.6.1.2: Comparison of both Automatic and Human Evaluation on Bing

Bing translation service's evaluation results show us the rates higher and closer to each other. Bing manage suffixes well when translation from English to Turkish

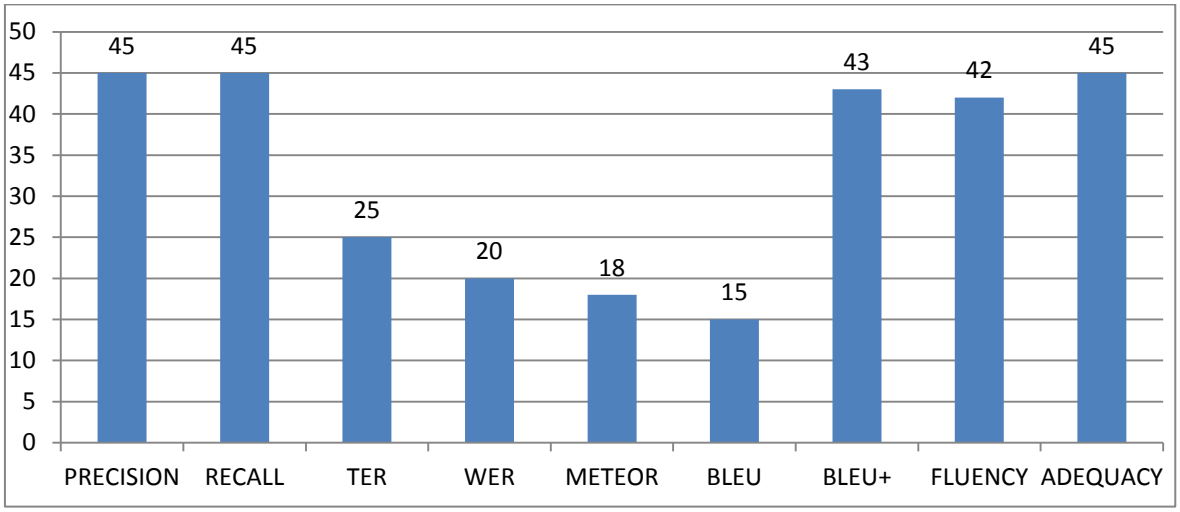


Figure 4.6.1.3: Comparison of both Automatic and Human Evaluation on Yandex

With higher Google but lower Bing automatic rates, evaluation results of Yandex. They are almost same both human and word detection rate metrics are Precision and Recall.

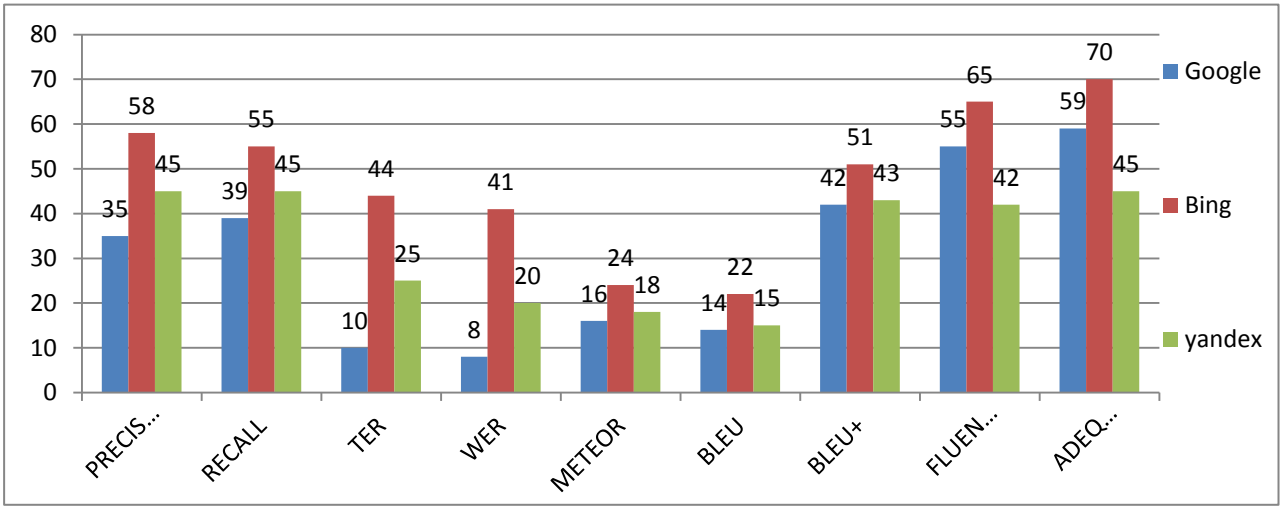


Figure 4.6.1.4: Comparison of Automatic and Human Evaluation

Randomly selected bilingual sentence corpus evaluation shows us that Bing translation service is better than Yandex and Yandex is better than Google in Turkish language comparison evaluation with Auto metrics. But human judgment tests are presenting that Google better than Yandex because of correct **synonym** selection then same root of related word or correct phrase selection of related word group. In both of state, Bing translate closed to Human approach is far from/ better than other services.

Precision rates are usually equal or smaller than recall and adequacy scores are generally bigger than fluency. Results demonstrated that entity of words scores is higher than

alignment of words success. Fluency and Adequacy rates are coming from expert judgments. And comparison of entire corpus and small subset of corpus to validate with human judgment exhibited that using **synonym** of word or **phrases** reason low auto metric score than human judgment score.

Every Adequacy and Fluency scores are coming from average score of 2 different human experts. Alignment of word as presentenced as TER, WER, Meteor and Bleu high Human judgment criteria, Adequacy and Fluency is high better correlation.

Table 4.6.1.2: Evaluation from English to Turkish Sample Similarity Score Table

#	Source-En	Reference-Tr	MT-Tr	Prec.	Recall	TER	WER	Meteor	Bleu	Bleu+	Adequacy	Fluency
1	The fuel cell system is simulated under different AC load conditions.	Yakıt pili sistemi farklı AC yük şartları altında simüle edilmiştir.	Yakıt hücresi sistemi farklı bir AC yük şartlarında simüle edilmektedir.	0.6	0.6	0.5	0.5	0.2	0.12	0.6	0.8	0.8
2	America's best known novelists, journalists, and editors attended a conference in New York last week.	Amerika'nın en ünlü romancıları, gazetecileri ve editörlerigeçen hafta New York'ta bir konferansa katıldılar.	Amerika'nın en tanınmış romancılar, gazeteciler, editörler ve geçen hafta New York'ta bir konferansa katıldı.	0.53	0.61	0.5	0.5	0.29	0.25	0.31	0.6	0.6
3	They decided to get married next month.	Gelecek ay evlenmeye karar verdiler.	Gelecek ay evlenmeye karar.	0.5	0.6	0.4	0.4	0.30	0.26	0.56	0.5	0.5
…												
32	Many people believe that an ulcer is caused by stress or spicy foods, but this is not the case.	Pek çok insan, ülserin, stres veya baharatlı gıdalar nedeniyle oluştuğuna inanır, ancak durum böyle değildir.	Birçok kişi bir ülser stres veya baharatlı gıdalar neden olduğuna inanıyoruz, ama bu durum böyle değil.	0.2	0.2	0.1	0.1	0.1	0.1	0.2	0.6	0.8

4.6.2 Human Evaluation over English Train Subset

This sub section shows the evaluation for English Sentence pairs translated from Turkish subset. Results are as follows:

Table 4.6.2.1: Comparatively Human and Automatic Evaluation of English Corpus

Set / Methods	Precision	Recall	TER	WER	Meteor	Bleu	Bleu+	Fluency	Adequacy
Google	60	58	45	38	30	32	43	61	68
Bing	55	53	36	27	26	26	45	62	63
Yandex	44	43	19	13	15	14	48	45	47

The Table 4.6.2.1 above and Figure 4.6.2.1 show us the evalaution comparison of online translation service over auto and manual metrics with random bilingual corpus which contain 32 english source texts and tranalated sentence from Turkish to English language.

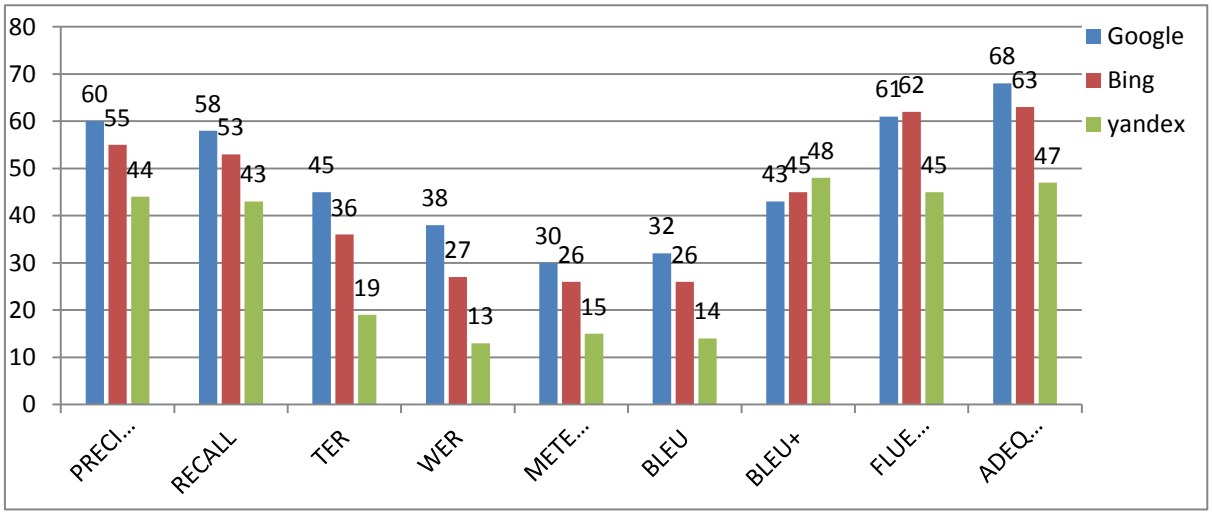


Figure 4.6.2.1: Comparatively Human and Automatic Evaluation of English Corpus

Results show that Google is better than others in Turkish to English translation generally. Then Bing and Yandex perform lower quality. And it can be say easily that parabolic type of metric aligns shows us the edge of metric columns are representing word detection rates and middle part is give evidence about word sequence.

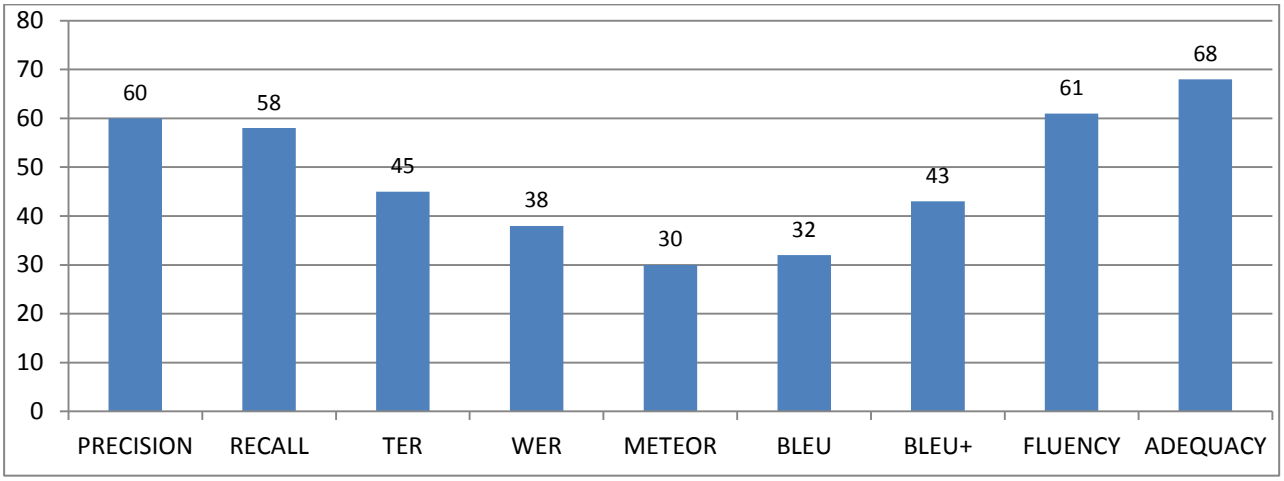


Figure 4.6.2.2: Evaluation Comparison by Metrics on Google for English Corpus

It is seen clearly that parabolic curve slope is more soft because of well design of word occurred in sentence. And an observation can be obtained from the figure that expert human approach is almost same rate with Precision and Recall from Turkish to English languages translation by Google.

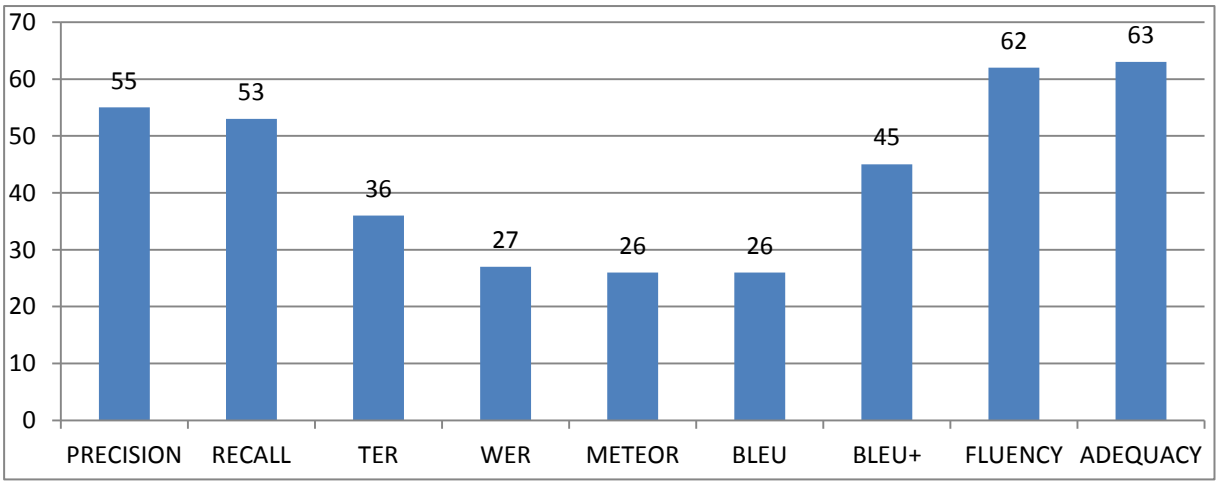


Figure 4.6.2.3: Evaluation Comparison by Metric on Bing for English Corpus

With a little bit lower similarity evaluation rates, Bing is on the second line after Google. Although bleu rates are lower than Google Bleu metric rate, Bleu+ metric gives greater rate for Bing translation service translation from Turkish to English language than Google. It shows that as in English to Turkish translation, on From Turkish to English translation Bing gives more fluent approach than Google.

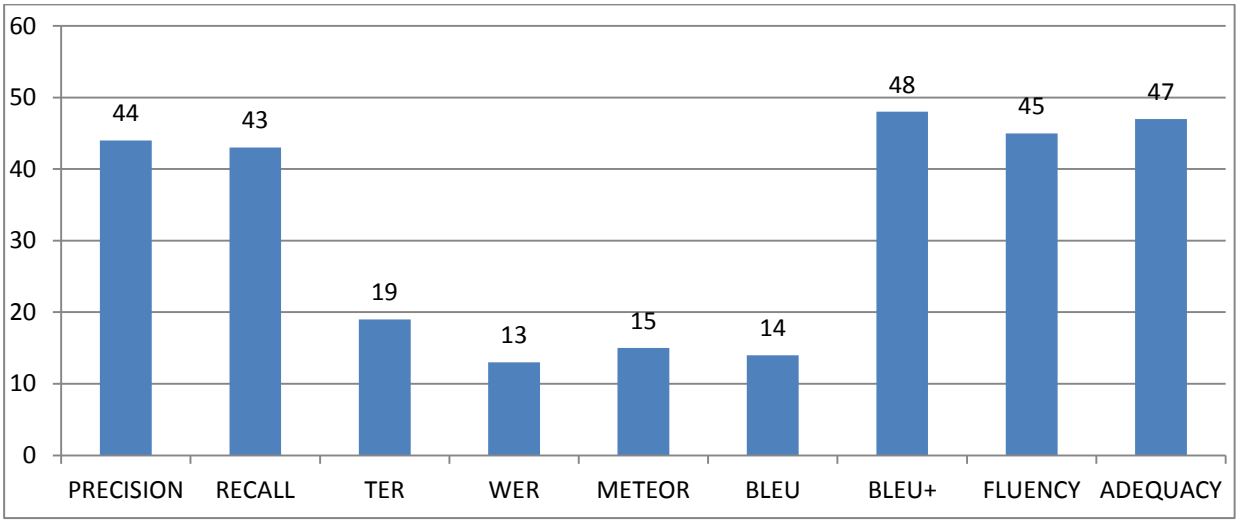


Figure 4.6.2.4: Evaluation Comparison by Metric over Yandex English Corpus

It is figured out from Yandex similarity evaluation results Figure 4.6.2.4 is that over English corpus already translated from Turkish corpus is Yandex lowest performs the lowest performance. However bleu+ rate on 48% rate, Human approach is only on the 45% with 47% rate of adequacy. So for three of online translation service, Adequacy rate is bigger than fluency. Human mind thinks that word location is exchangeable and it is redesign in human brain to understand a text. But in first impression wrong word sequence is affect less meaning about text.

Table 4.6.2.2: Sample Sentences and Their Evaluation Metric Scores

#	Source-Tr	Reference-En	MT-En	P	R	T	W	M	B	B+	Adequacy	Fluency
1	Yakıt pili sistemi farklı AC yük şartları altında simüle edilmiştir.	The fuel cell system is simulated under different AC load conditions.	The fuel cell system was simulated under different AC load conditions.	0.91	0.91	0.91	0.91	0.55	0.73	0.70	0.40	0.60
2	Amerika'nın en ünlü romancıları, gazetecileri ve editörlerigeçen hafta New York'ta bir konferansa katıldılar.	America's best known novelists, journalists, and editors attended a conference in New York last week.	America's most famous novelists, journalists and editörleri geç attended a conference in New York last week.	0.73	0.73	0.73	0.73	0.34	0.55	0.50	0.40	0.80
3	Gelecek ay evlenmeye karar verdiler.	They decided to get married next month.	They decided to get married next month.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
…	…	..	..									
32	Pek çok insan, ülserin, stres veya baharatlı gıdalar nedeniyle oluştuğuna inanır, ancak durum böyle değildir.	Many people believe that an ulcer is caused by stress or spicy foods, but this is not the case.	Many people, ulcer, caused by stress or spicy foods believes, however, is not the case.	0.67	0.53	0.53	0.53	0.30	0.44	0.32	0.40	0.60

When precision score is smaller than recall, it means that while number of matching word is same, word number of candidate sentence is smaller than references. So it affects bleu score via decreasing brevity penalty multiplier.

It ıs understood from average measurement results that bad translation samples with a little bit good translation sample decreased average score of selected corpora.

CHAPTER 5

COMPARISON OF THE FINDINGS AND DISCUSSION

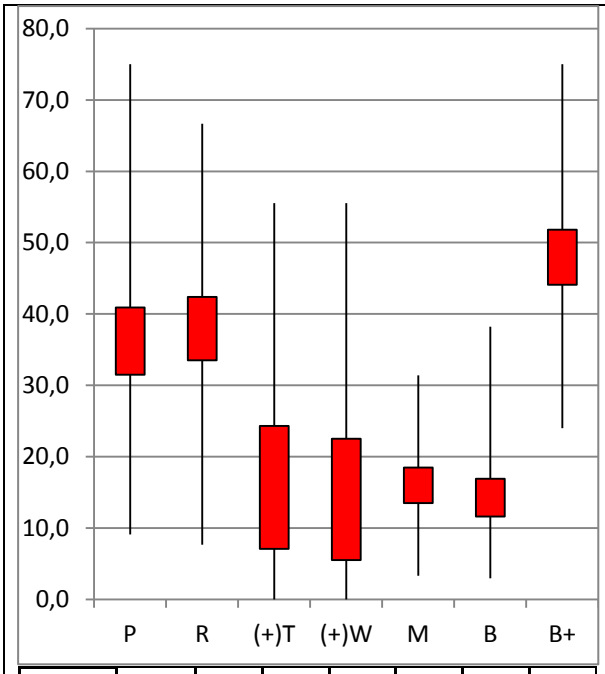
5.1 Comparatively Evaluations

The following figures and tables show the comparison of web services in terms of output quality successful by best/high score automatic method best correlated with expert mind according to sentence type from Turkish source to English translation. Table below gives us a general approach. Generally, Bleu+ metric gives us bigger score than bleu because bleu+ calculate suffix near word similarity and checking synonyms from the list. So, for instance, “book” and “books” are different words at surface based but they are almost same. So Bleu+ get us obtain that much closer evaluation score better correlated with human translation. Google English to Turkish translation evaluation was assessed over 50 random selected bilingual simple Turkish sentences pairs translated from English source text by human and machine. **Bleu+, TER, Meteor, Recall** metrics are more meaningful for comparison of each text comparatively. But in many figures and tables **Precision** and **WER** metrics are also used in tests comparatively.

The table below presents highest rates over sentence structure and translation web services in Turkish corpora. As seen on the figures that online machine translation services and best indicator methods are determined with services ability over sentence structures.

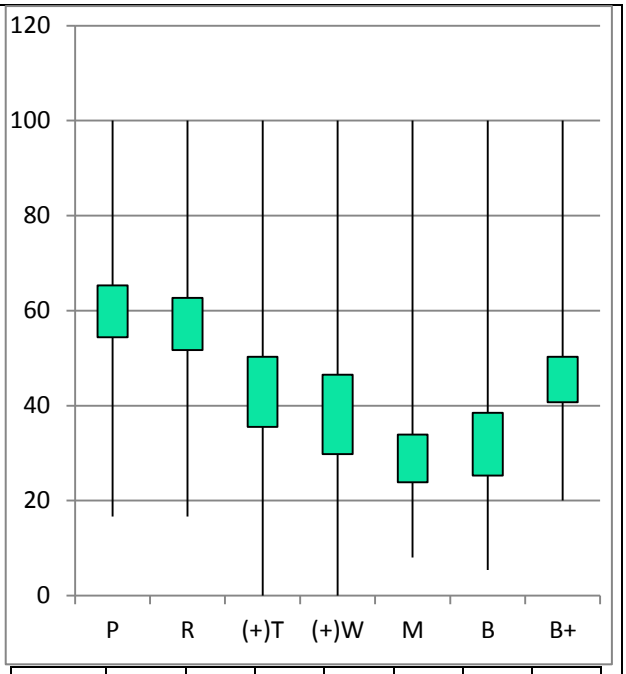
Google English to Turkish evaluation was performed by Asiya Evaluation tool over 50 random selected bilingual sentences pairs translated from English to Turkish languages.

Table 5.1.1: Min, Max, Confidence Interval Bottom and Top level Scores of Evalaution Rates for Simple Turkish (A) and English (B) Corpus on Google



	P	R	T	W	M	B	B+
Lower Bound	31.5	33.5	7.1	5.5	13.5	11.6	44.1
Max	75.0	66.6	55.5	55.5	31.4	38.2	75.0
min	9.0	7.6	0.0	0.0	3.3	2.9	24.0
Upper Bound	40.9	42.4	24.3	22.5	18.5	16.9	51.8

(A)

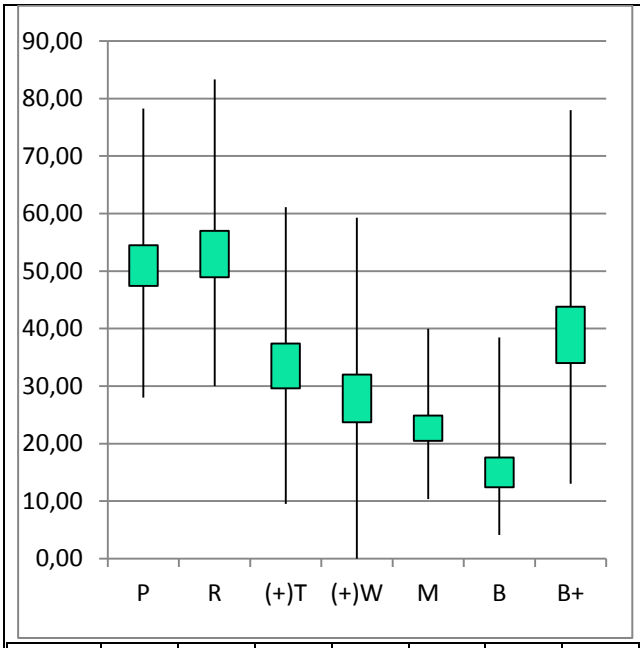


	P	R	T	W	M	B	B+
Upper Bound	65.3	62.7	50.3	46.5	33.9	38.5	50.3
Max	100	100	100	100	100	100	100
min	16.6	16.6	0	0	8.02	5.42	20
Lower Bound	54.4	51.7	35.5	29.8	23.9	25.3	40.7

(B)

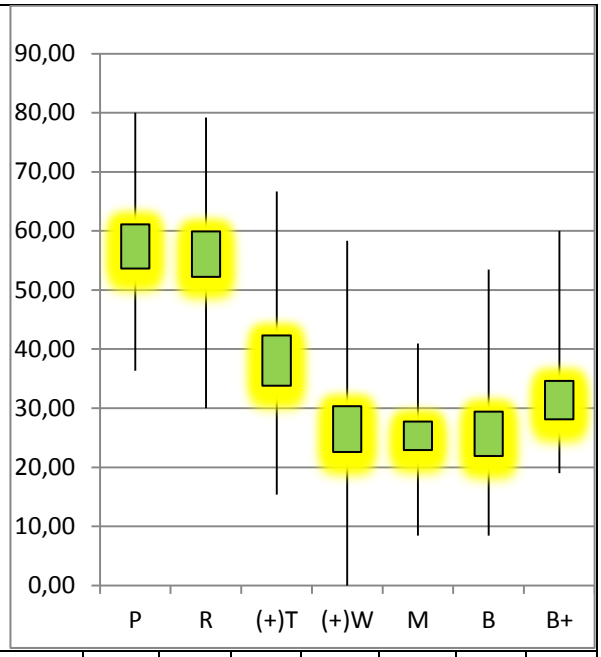
According to Table 5.1.1 colorful parts are representing coverage range of verification test results. This range means that standard deviation is greater than other metrics. Small ranges are representing the more stable score.

Table 5.1.2: Min, Max, Confidence Interval Bottom and Top level Scores of Evalaution Rates for Compound Turkish (A) and English (B) Corpus on Google



	P	R	T	W	M	B	B+
Lower Bound	47.4	48.9	29.6	23.7	20.5	12.4	34.0
Upper Bound	54.5	57.0	37.4	32.0	24.9	17.6	43.8
min	28.0	30.0	9.5	0.0	10.3	4.1	13.0
Max	78.2	83.3	61.1	59.2	39.9	38.4	78.0

A)

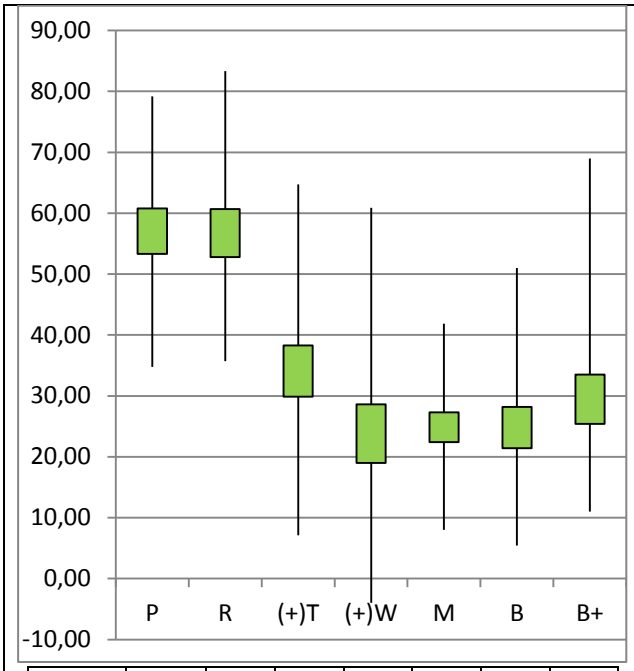


	P	R	T	W	M	B	B+
Lower Bound	53.6	52.2	33.8	22.6	22.9	21.9	28.1
Upper Bound	61.1	59.9	42.3	30.3	27.7	29.4	34.6
min	36.3	30.0	15.3	0.0	8.4	8.45	19.0
Max	80.0	79.1	66.6	58.3	40.9	53.4	60.0

B)

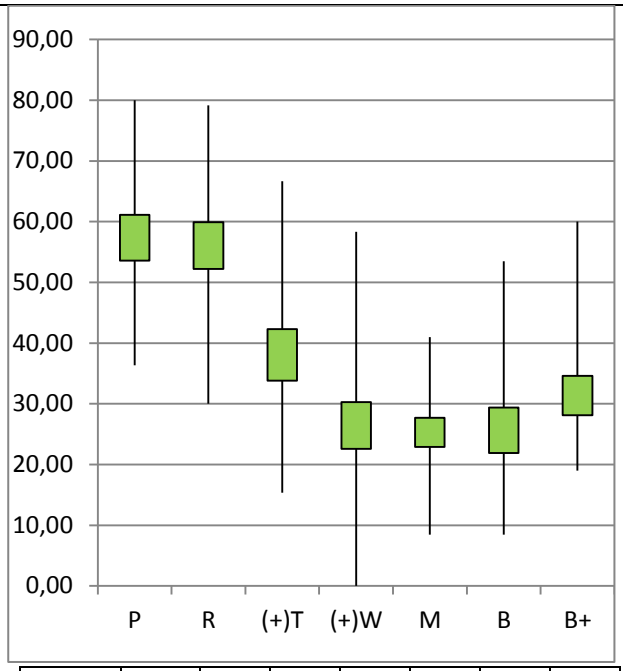
Translation quality evaluation was performed from English to Turkish on the left side on compound sentences and evaluation results on the right side were figured from Turkish to English by Google. With the same intersected matching words, recall rates are greater than precision rates on from English to Turkish evaluation.

Table 5.1.3: Min, Max, Confidence Interval Bottom and Top level Scores of Evalaution Rates for Complex Turkish (A) and English (B) Corpus on Google



	P	R	T	W	M	B	B+
Lower Bound	53.3	52.8	29.9	19.0	22.4	21.4	25.4
Upper Bound	60.8	60.7	38.3	28.6	27.3	28.2	33.5
min	34.7	35.7	7.1	-4.0	8.0	5.4	11.0
Max	79.1	83.3	64.7	60.8	41.8	51.0	69.0

(A)



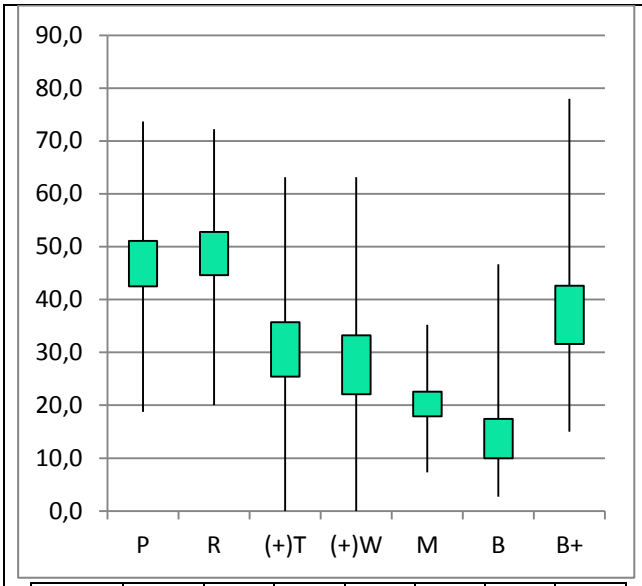
	P	R	T	W	M	B	B+
Lower Bound	53.6	52.2	33.8	22.6	22.9	21.9	28.1
Upper Bound	61.1	59.9	42.3	30.3	27.7	29.4	34.6
min	36.3	30.0	15.3	0.0	8.4	8.4	19.0
Max	80.0	79.1	66.6	58.3	40.9	53.4	60.0

(B)

Simple and complex sentences are actually better translation with Bing form English to Turkish but compound sentences are well translated with Google and Yandex than Bing service. Translations from Turkish to English are usually more meaningful than from English to Turkish since exact statistical results, Meteor and Bleu scores are more acceptable. But translations from English to Turkish are more compatible by Bleu+ method which is closed to Precision and Recall. If Recall scores are bigger than Precision, Precision based Bleu scores are smaller than Recall based Meteor. Bleu+ examines suffixes with word root/stem and it uses synonym comparison list.

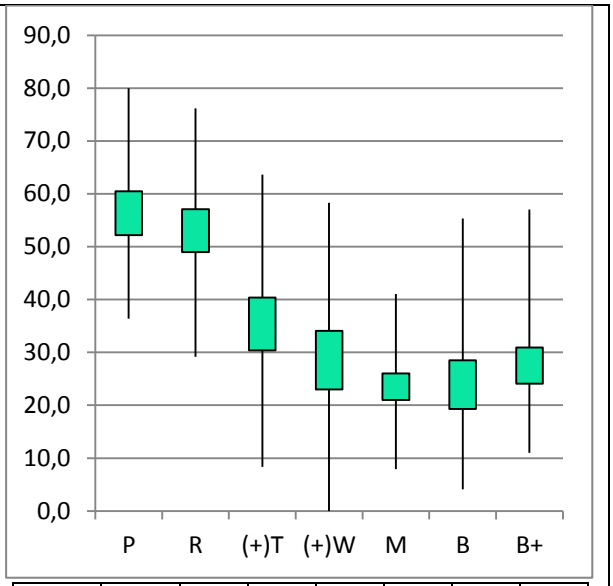
.

Table 5.1.4: Min, Max, Confidence Interval Bottom and Top level Scores of Evalaution Rates for Compound-Complex Turkish (A) and English (B) Corpus on Google



	P	R	T	W	M	B	B+
Lower Bound	42.5	44.6	25.4	22.1	17.9	10.0	31.6
Upper Bound	51.1	52.8	35.7	33.2	22.6	17.4	42.6
min	18.8	20.0	0.0	0.0	7.3	2.7	15.0
Max	73.7	72.2	63.2	63.2	35.3	46.7	78.0

(A)

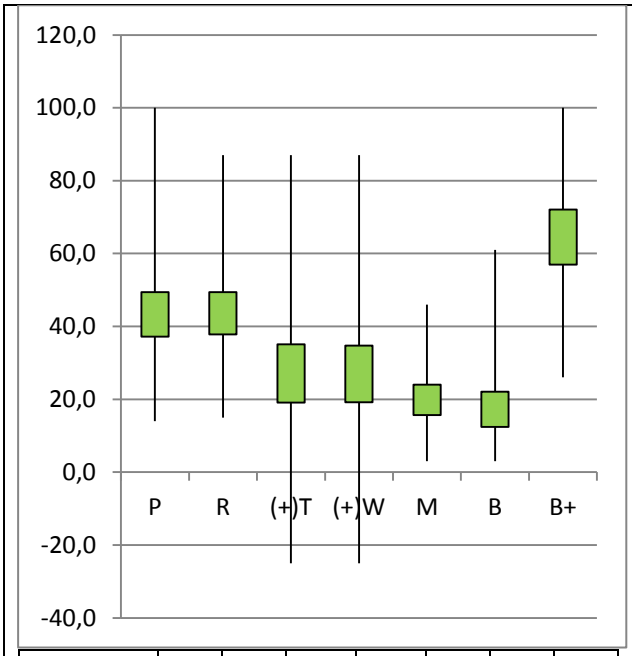


	P	R	T	W	M	B	B+
Lower Bound	52.2	49.0	30.4	23.0	21.0	19.3	24.1
Upper Bound	60.5	57.1	40.4	34.1	26.0	28.5	30.9
min	36.4	29.2	8.3	0.0	7.9	4.1	11.0
Max	80.0	76.2	63.6	58.3	41.1	55.3	57.0

(B)

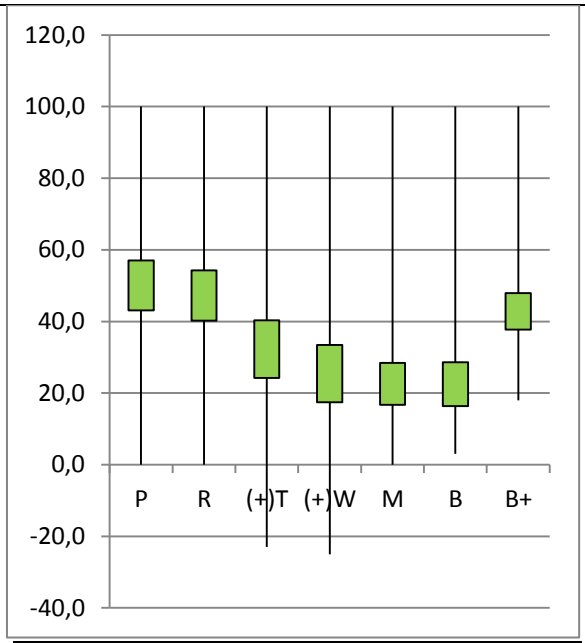
Google service evaluation from English to Turkish language on compound-complex sentence structure with 40 sentences was performed on the left side of in Table 5.1.4 and evaluation from Turkish to English on the right side. It is clearly seen that Meteor metric confidence intervals are very tight on the both side. This means that almost every sentence in range of 5 (from 17 to 26 or from 21 to 26). Because, many words are occurred in Turkish sentences have suffixes, Meteor metric gives a little better score on evaluation from Turkish to English since there are more exactly matched words in English.

Table 5.1.5: Min, Max, Confidence Interval Bottom and Top level Scores of Evalaution Rates for Simple Turkish (A) and English (B) Corpus on Bing



	P	R	T	W	M	B	B+
Lower Bound	37.2	37.8	19.1	19.2	15.7	12.4	57.0
Upper Bound	49.4	49.4	35.1	34.7	24.0	22.1	72.1
min	14.0	15.0	-25.0	-25.0	3.0	3.0	26.0
Max	100	87.0	87.0	87.0	46.0	61.0	100

(A)

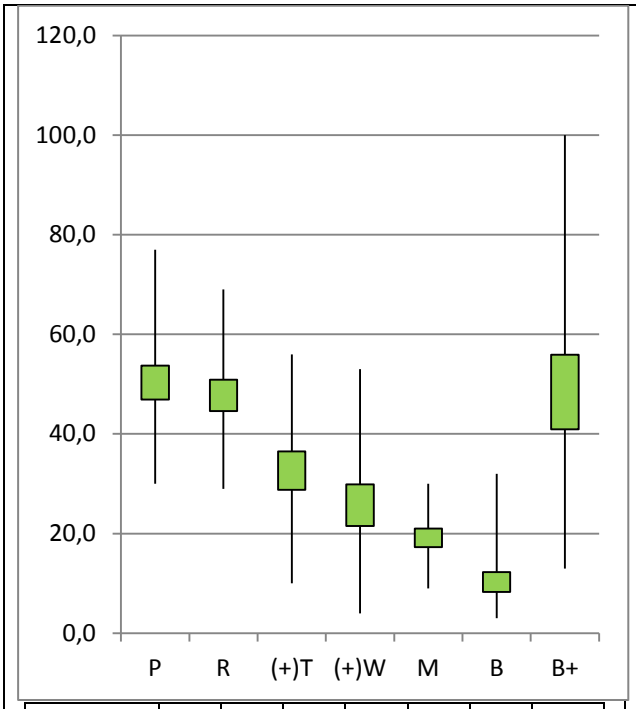


Metric	P	R	T	W	M	B	B+
Lower Bound	43.1	40.2	24.2	17.4	16.7	16.4	37.7
Upper Bound	57.0	54.3	40.3	33.4	28.4	28.6	47.9
min	0.0	0.0	-23.0	-25.0	0.0	3.0	18.0
Max	100	100	100	100	100	100	100

(B)

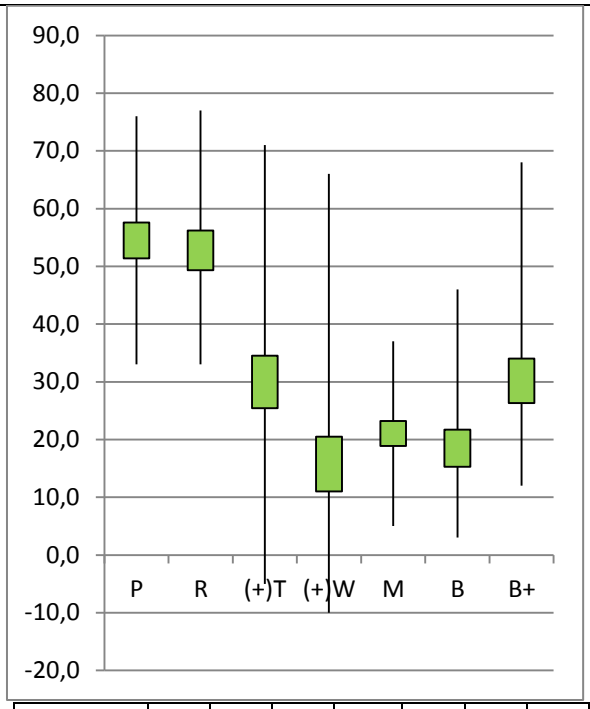
It is obviously seen that Bleu and Bleu+ differentiation is greater in English to Turkish evaluation of Bing translation service with simple sentences than Turkish to English. Meteor and Bleu metrics give almost same rates. This means that Meteor may be used insist of Bleu or vice versa. TER and WER metrics give some evidence about word location and number of word. So reference-candidate sentence lengths in terms of word are showing diversity. Number of word in both reference and candidate translation sentence are not equal one to one.

Table 5.1.6: Min, Max, Confidence Interval Bottom and Top level Scores of Evalaution Rates for Complex Turkish (A) and English (B) Corpus on Bing



Metric	P	R	T	W	M	B	B+
Lower Bound	46.9	44.6	28.8	21.5	17.3	8.3	40.9
Upper Bound	53.7	50.9	36.5	29.9	21.0	12.3	55.9
min	30.0	29.0	10.0	4.0	9.0	3.0	13.0
Max	77.0	69.0	56.0	53.0	30.0	32.0	100.0

(A)

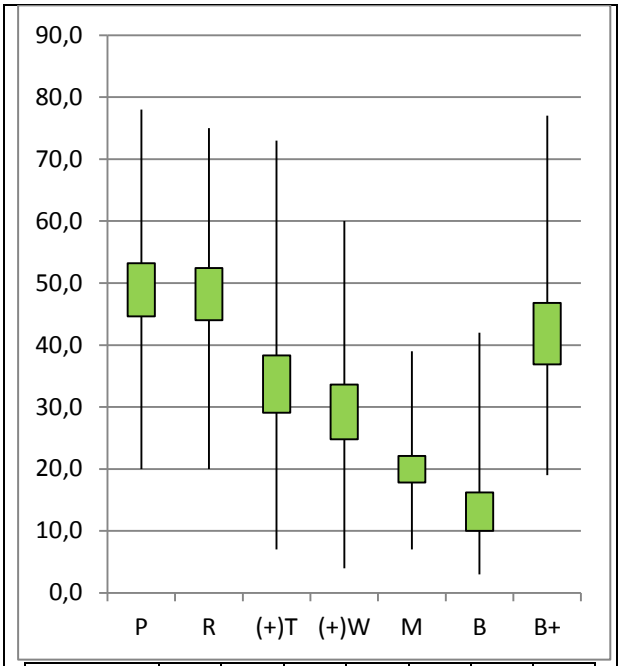


Metric	P	R	T	W	M	B	B+
Lower Bound	51.4	49.3	25.4	11.0	18.9	15.3	26.3
Upper Bound	57.6	56.2	34.5	20.5	23.2	21.7	34.0
min	33.0	33.0	-5.0	-10.0	5.0	3.0	12.0
Max	76.0	77.0	71.0	66.0	37.0	46.0	68.0

(B)

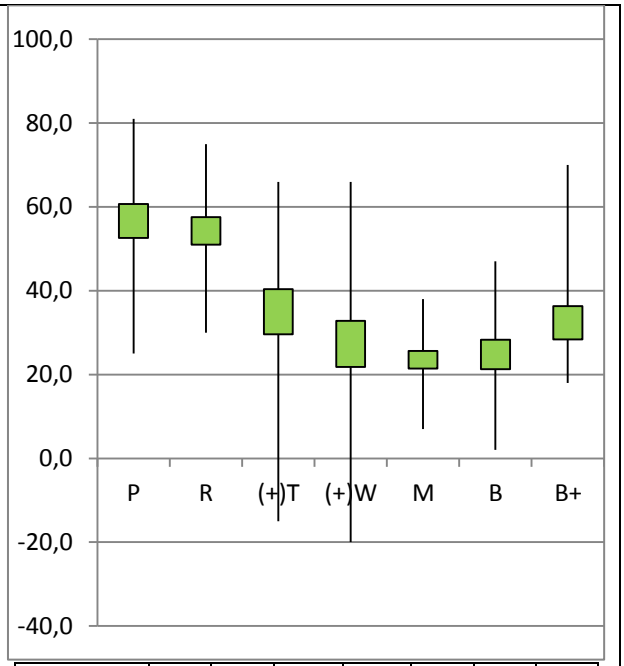
For complex sentence evaluation, Bing test results from English to Turkish show lower rates since there are less exactly matching words. It may be caused because of wrong word synonym selections.

Table 5.1.7: Min, Max, Confidence Interval Bottom and Top level Scores of Evalaution Rates for Compound Turkish (A) and English (B) Corpus on Bing



Metric	P	R	T	W	M	B	B+
Lower Bound	44.6	44.0	29.1	24.8	17.8	10.0	36.9
Upper Bound	53.2	52.4	38.3	33.6	22.1	16.2	46.8
min	20.0	20.0	7.0	4.0	7.0	3.0	19.0
Max	78.0	75.0	73.0	60.0	39.0	42.0	77.0

(A)

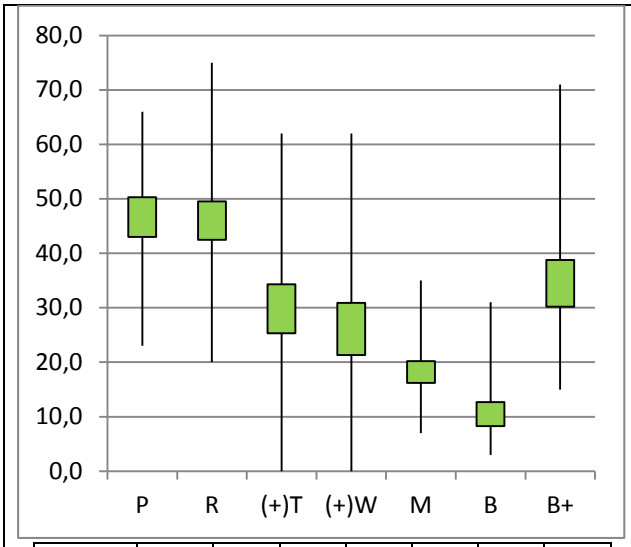


Metric	P	R	T	W	M	B	B+
Lower Bound	52.6	51.0	29.6	21.8	21.4	21.3	28.4
Upper Bound	60.7	57.6	40.4	32.8	25.6	28.3	36.3
min	25.0	30.0	-15.0	-20.0	7.0	2.0	18.0
Max	81.0	75.0	66.0	66.0	38.0	47.0	70.0

(B)

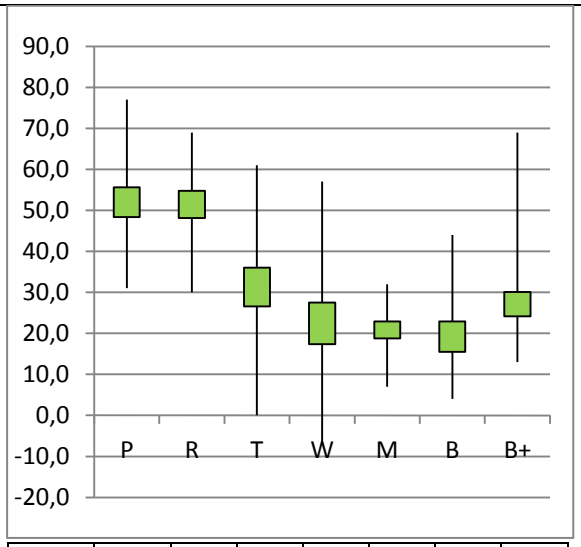
Metrics give hints about compound sentence evaluation on Bing translation service. It can be impressed that from Turkish to English translation on Bing translation service can provide exactly matched words better than English to Turkish translation. The reason for this is because of wrong synonym or parallel meaning of word or word groups with lack of suffixes.

Table 5.1.8: Min, Max, Confidence Interval Bottom and Top level Scores of Evalaution Rates for Compound-Complex Turkish (A) and English (B) Corpus on Bing



	P	R	T	W	M	B	B+
Lower Bound	43.0	42.5	25.3	21.3	16.2	8.3	30.2
Upper Bound	50.3	49.5	34.3	30.9	20.2	12.7	38.8
min	23.0	20.0	0.0	0.0	7.0	3.0	15.0
Max	66.0	75.0	62.0	62.0	35.0	31.0	71.0

(A)

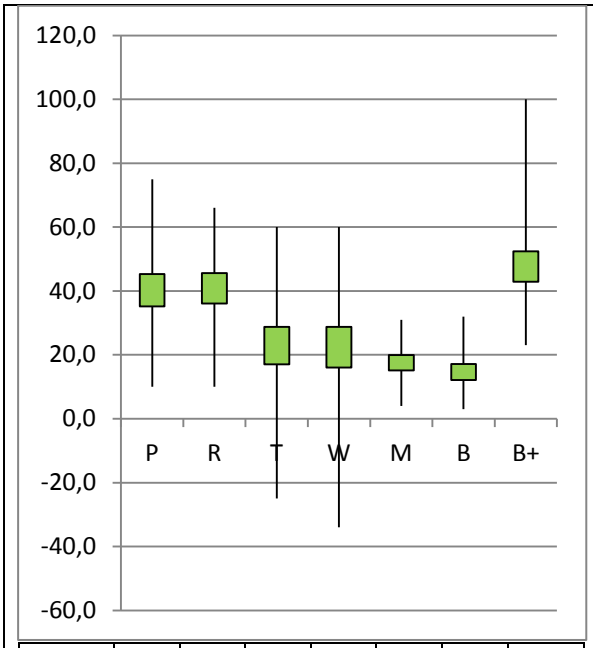


	P	R	T	W	M	B	B+
Lower Bound	48.4	48.1	26.6	17.4	18.8	15.5	24.2
Upper Bound	55.6	54.8	36.0	27.5	22.9	22.9	30.1
min	31.0	30.0	0.0	-7.0	7.0	4.0	13.0
Max	77.0	69.0	61.0	57.0	32.0	44.0	69.0

(B)

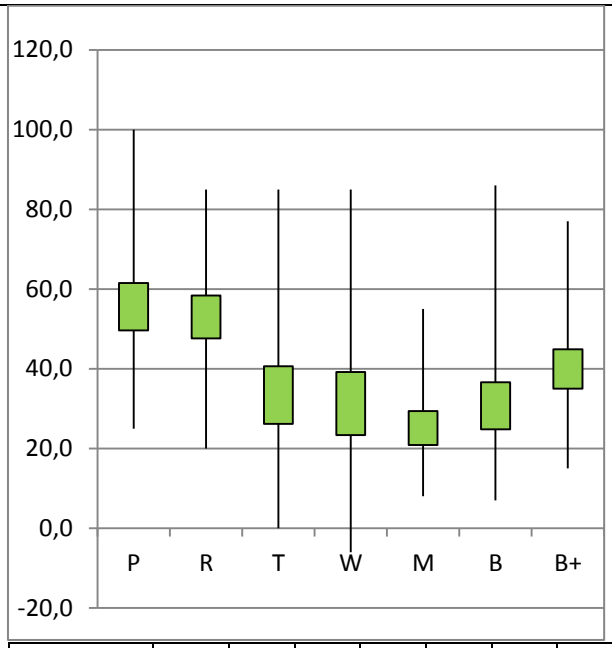
Lastly for Bing translation service, for complex-compound evaluation test results by metrics show almost same results about rates of metrics each other. It is seen by means of Bleu-Bleu+ ratio that there are some words same root different suffix. So average Bleu+ rate give information about word status of bilingual sentences that Turkish text specific produced metric Bleu+ is also can be used for English texts.

Table 5.1.9: Min, Max, Confidence Interval Bottom and Top level Scores of Evalaution Rates for Simple Turkish (A) and English (B) Corpus on Yandex



Metric	P	R	T	W	M	B	B+
Lower Bound	35.2	36.1	17.0	16.0	15.1	12.1	42.9
Upper Bound	45.3	45.6	28.8	28.8	19.9	17.1	52.4
min	10.0	10.0	-25.0	-34.0	4.0	3.0	23.0
Max	75.0	66.0	60.0	60.0	31.0	32.0	100.0

(A)

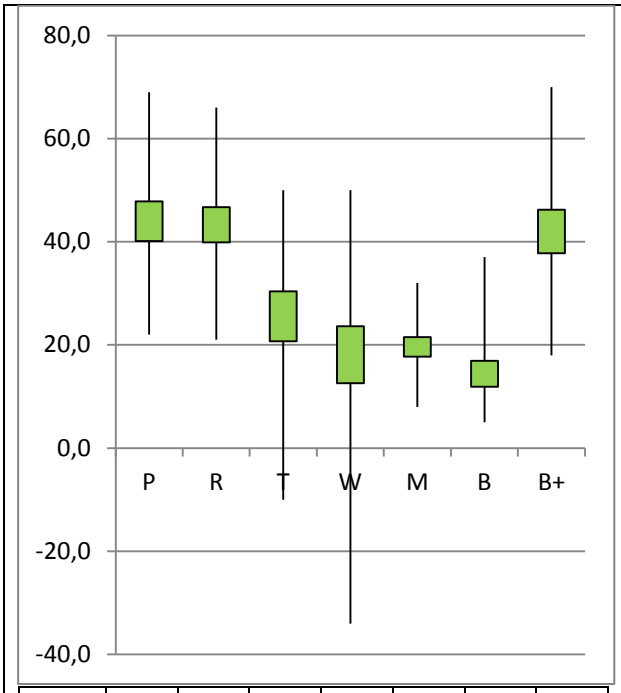


	P	R	T	W	M	B	B+
Lower Bound	49.6	47.6	26.2	23.4	20.9	24.8	35.0
Max	61.5	58.4	40.6	39.2	29.4	36.6	44.9
min	25.0	20.0	0.0	-6.0	8.0	7.0	15.0
Upper Bound	100.0	85.0	85.0	85.0	55.0	86.0	77.0

(B)

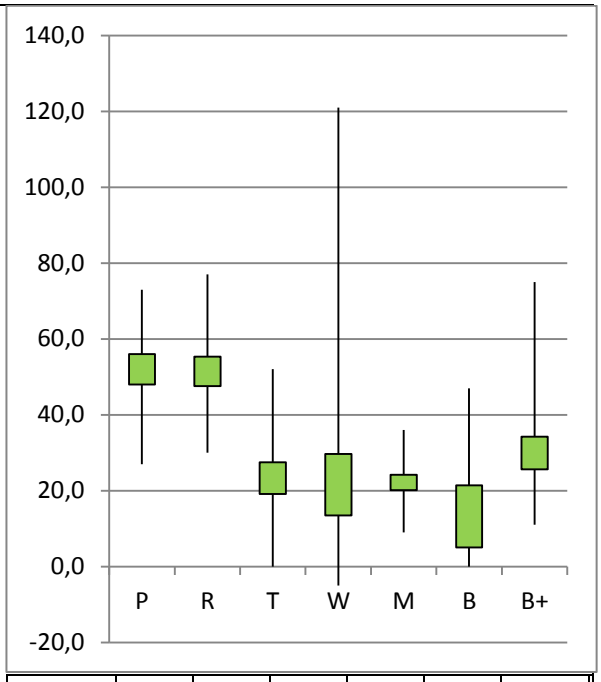
It is surprisingly a new acquisition that directly exact word matching metrics, Precision and Recall, are giving lower similarity rate than Bleu+ on simple sentence translation. It shows that many words can be translate almost same on root level exclude suffix. Small suffix diversity was ignored by the Bleu+ and alignment of word also almost perfect in English to Turkish. Moreover, there are no distinctive differences from previous evaluation test on Turkish to English translation. These results are required post-editing process during translation.

Table 5.1.10: Min, Max, Confidence Interval Bottom and Top level Scores of Evalaution Rates for Complex Turkish (A) and English (B) Corpus on Yandex



	P	R	T	W	M	B	B+
Lower Bound	40.1	39.9	20.7	12.6	17.7	11.9	37.8
Max	47.8	46.7	30.4	23.6	21.5	16.9	46.2
min	22.0	21.0	-10.0	-34.0	8.0	5.0	18.0
Upper Bound	69.0	66.0	50.0	50.0	32.0	37.0	70.0

(A)

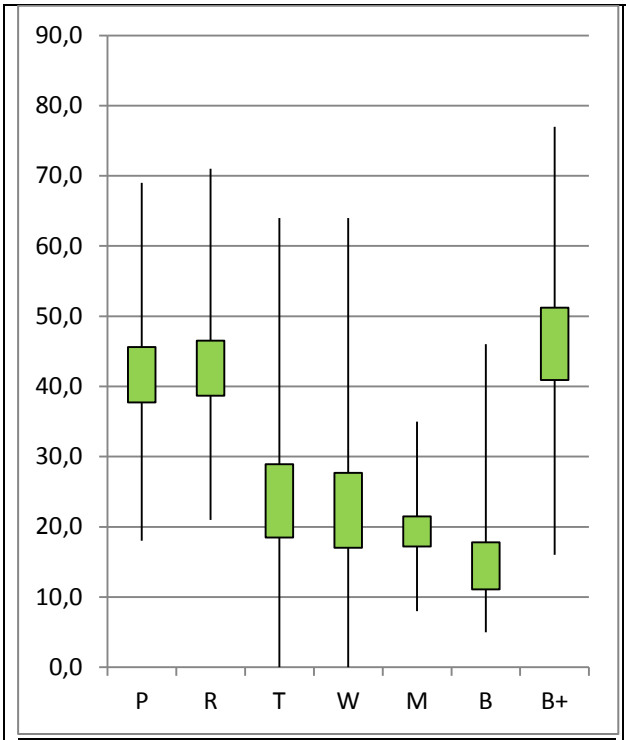


Metric	P	R	T	W	M	B
Lower Bound	48.0	47.6	19.1	13.5	20.1	5.0
Max	56.0	55.3	27.5	29.7	24.2	21.4
min	27.0	30.0	0.0	-5.0	9.0	0.0
Upper Bound	73.0	77.0	52.0	121	36.0	47.0

(B)

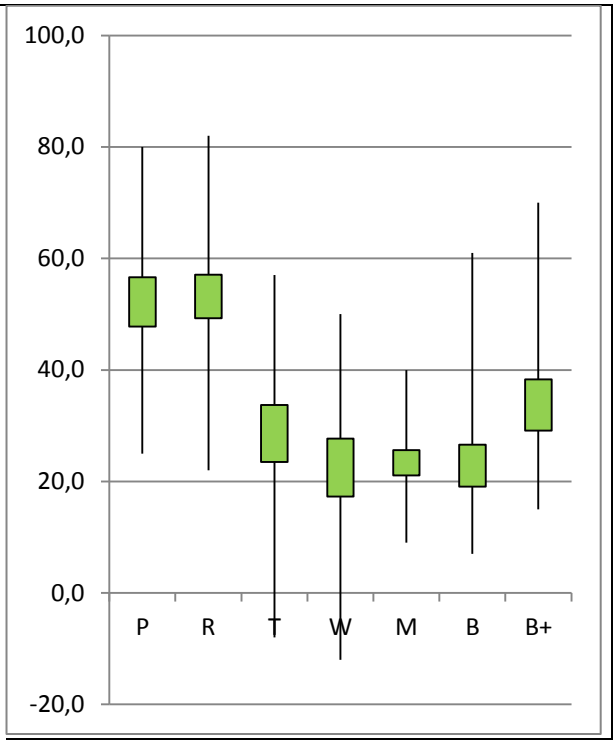
It is seen clearly again that there is no brevity penalty factor inside WER and TER metric formula on complex sentence translation. So since various numbers of words in reference and candidate translation, WER and WER might give scores under zero or over 100. In addition, it is illustrated that Precision, Recall and Bleu+ similarity rates are almost same. It is caused due to high same root different suffix and well-aligned word structure.

Table 5.1.11: Min, Max, Confidence Interval Bottom and Top level Scores of Evalaution Rates for Compound Turkish (A) and English (B) Corpus on Yandex



	P	R	T	W	M	B	B+
Lower Bound	37.7	38.7	18.5	17.0	17.2	11.1	40.9
Max	45.6	46.5	28.9	27.7	21.5	17.8	51.2
min	18.0	21.0	0.0	0.0	8.0	5.0	16.0
Upper Bound	69.0	71.0	64.0	64.0	35.0	46.0	77.0

(A)

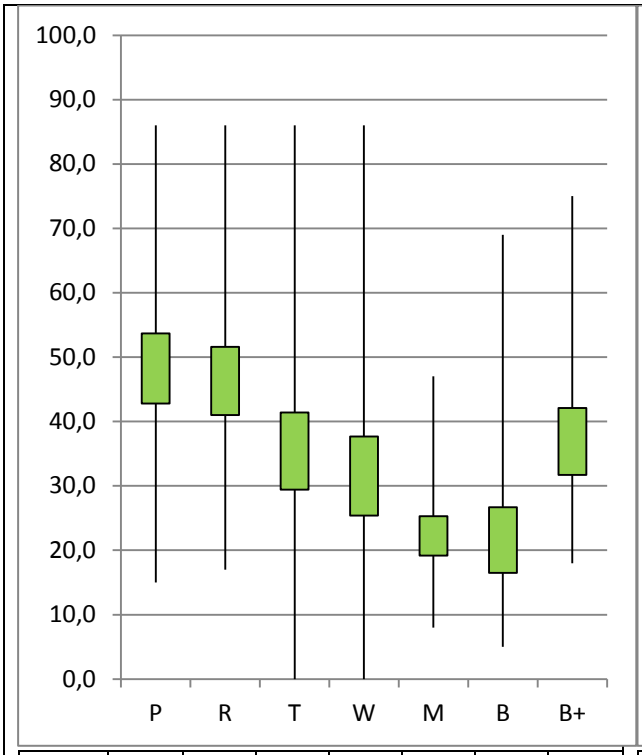


	P	R	T	W	M	B	B+
Lower Bound	47.8	49.3	23.5	17.3	21.1	19.1	29.1
Max	56.6	57.1	33.7	27.7	25.6	26.6	38.3
min	25.0	22.0	-8.0	-12.0	9.0	7.0	15.0
Upper Bound	80.0	82.0	57.0	50.0	40.0	61.0	70.0

(B)

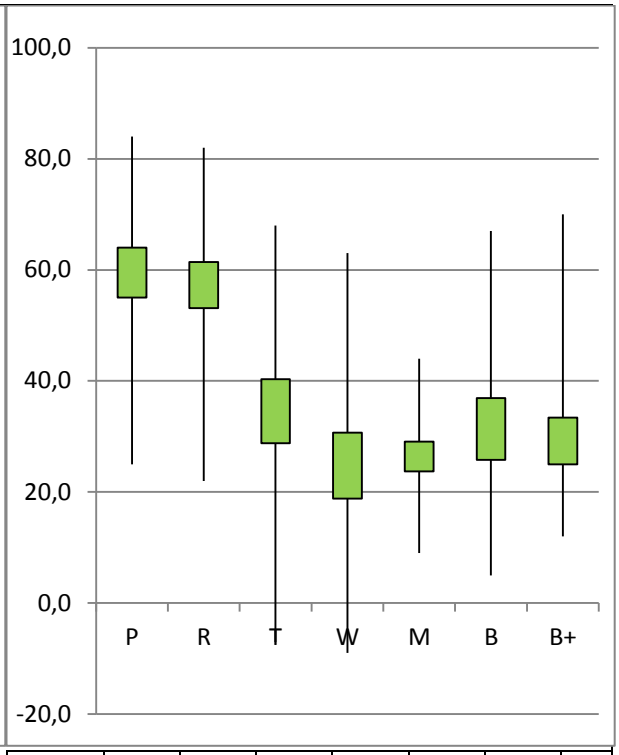
As seen in previous Yandex Evaluation from English to Turkish translation on simple and complex sentences, exact word and word root matching comparison by using with Precision/Recall and Bleu+ show us that Precision and Recall couldn't determine word root so they cannot reflect human approach about finding root and suffixes of word and also alignment of them.

Table 5.1.12: Min, Max, Confidence Interval Bottom and Top level Scores of Evalaution Rates for Compound-Complex Turkish (A) and English (B) Corpus on Yandex



	P	R	T	W	M	B	B+
Lower Bound	42.8	41.0	29.4	25.4	19.2	16.5	31.7
Max	53.7	51.6	41.4	37.7	25.3	26.7	42.1
min	15.0	17.0	0.0	0.0	8.0	5.0	18.0
Upper Bound	86.0	86.0	86.0	86.0	47.0	69.0	75.0

(A)



	P	R	T	W	M	B	B+
Lower Bound	55.0	53.1	28.8	18.8	23.7	25.8	25.0
Max	64.0	61.4	40.3	30.7	29.1	36.9	33.4
min	25.0	22.0	-7.0	-9.0	9.0	5.0	12.0
Upper Bound	84.0	82.0	68.0	63.0	44.0	67.0	70.0

(B)

It must be impressed that there is an exception about Bleu+. Generally Bleu+ rates must be equal or greater than Bleu metric rate but on evaluation of Yandex translation service on compound-complex sentences there is a contradiction. It might be caused due to punctuation or Bleu+ toll error.

Metric evaluation results with average scores vs. Sentence structure evaluation comparison over 3 different bilingual dataset are displayed on the following figure:

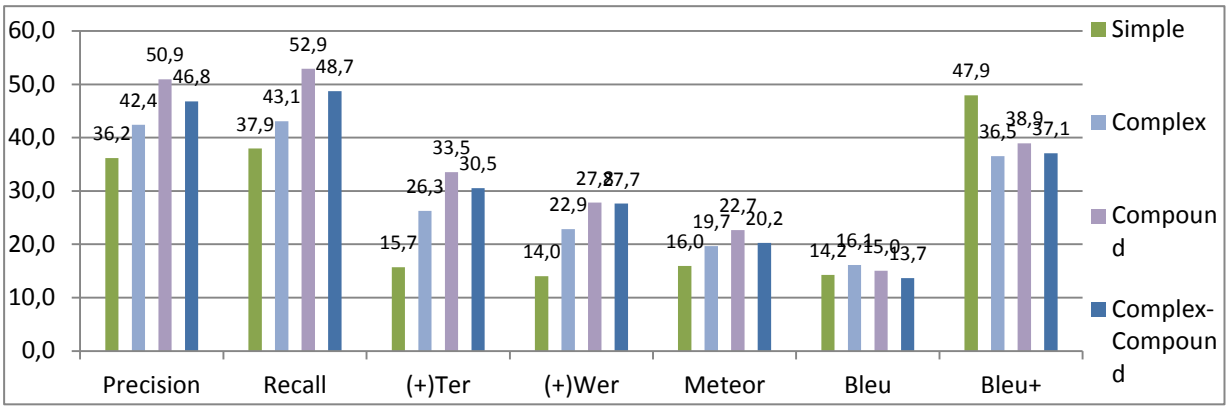


Figure 5.1.1: Average Scores of Evaluation Rates for All Structures of Turkish Sentences on Google

It is seen that translation of compound sentence in the bilingual corpus evaluated by auto metrics is with more quality that each other. But in term of Bleu+ metric simple translation is better on Google from English to Turkish.

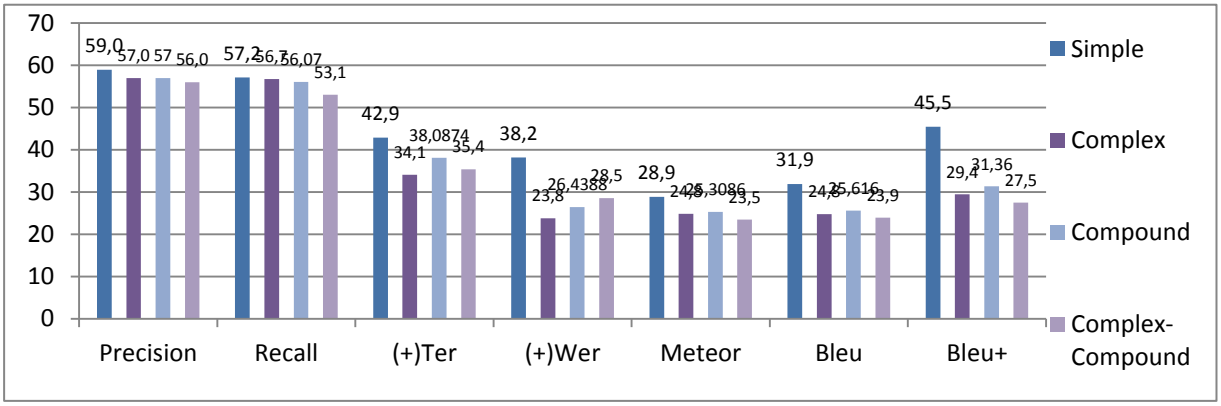


Figure 5.1.2: Average Scores of Evaluation Rates for All Structures of English Sentences on Google

Excluding Bleu+ metric, almost all metrics shows that Google can translate all sentence structure from Turkish to English almost same rate, but simple sentence can be translate better than others.

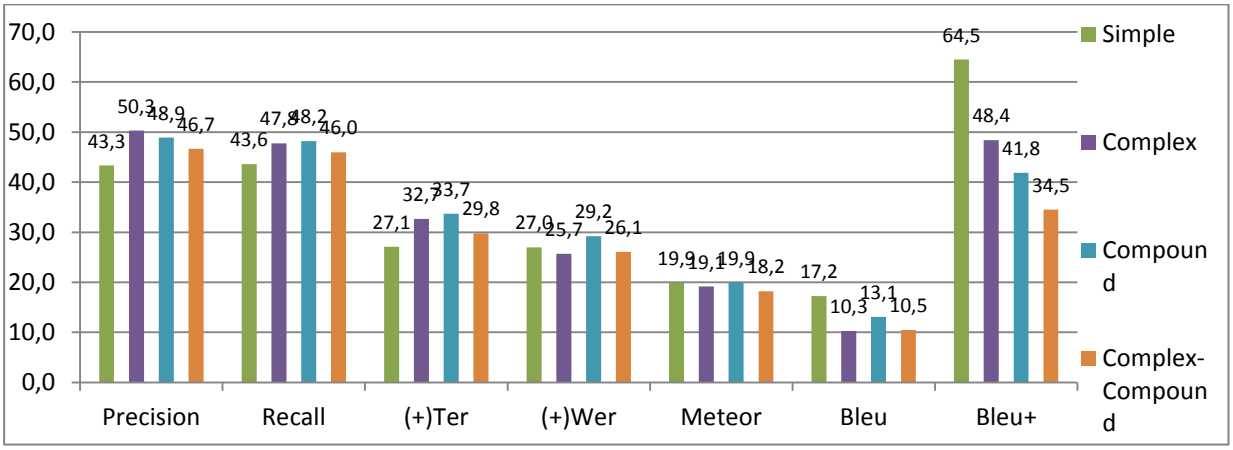


Figure 5.1.3: Average Scores of Evaluation Rates for All Structures of Turkish Sentences on Bing

Although complex sentence can be translated on word matching level, word alignment is better on compound sentence structure by the Bing translation service from English to Turkish. All metrics indicate that complex-compound sentences are translated by Bing with lower rates.

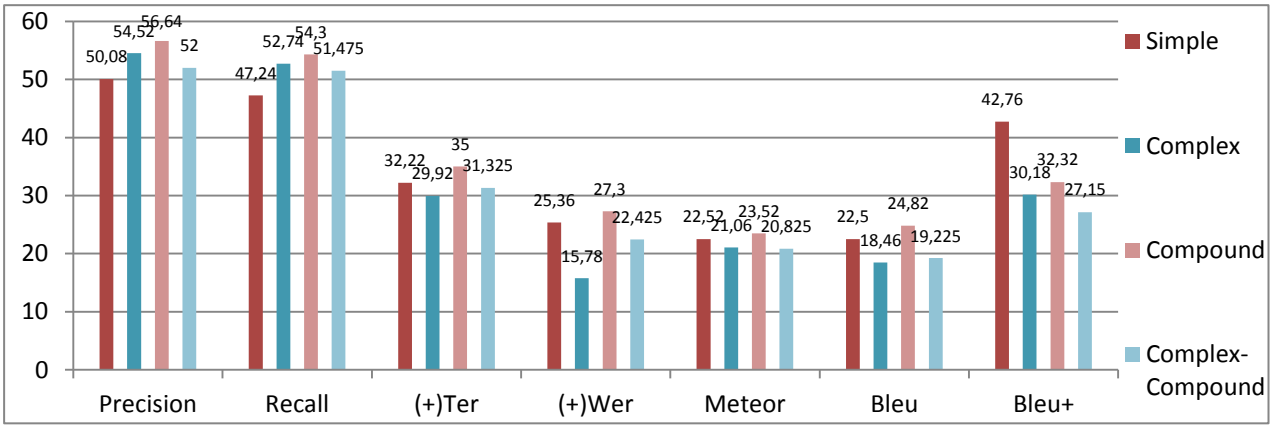


Figure 5.1.4: Average Scores of Evaluation Rates for All Structures of English Sentences on Bing

It must be impressed that compound sentence can be translated better than other type of sentence by Bing from Turkish to English.

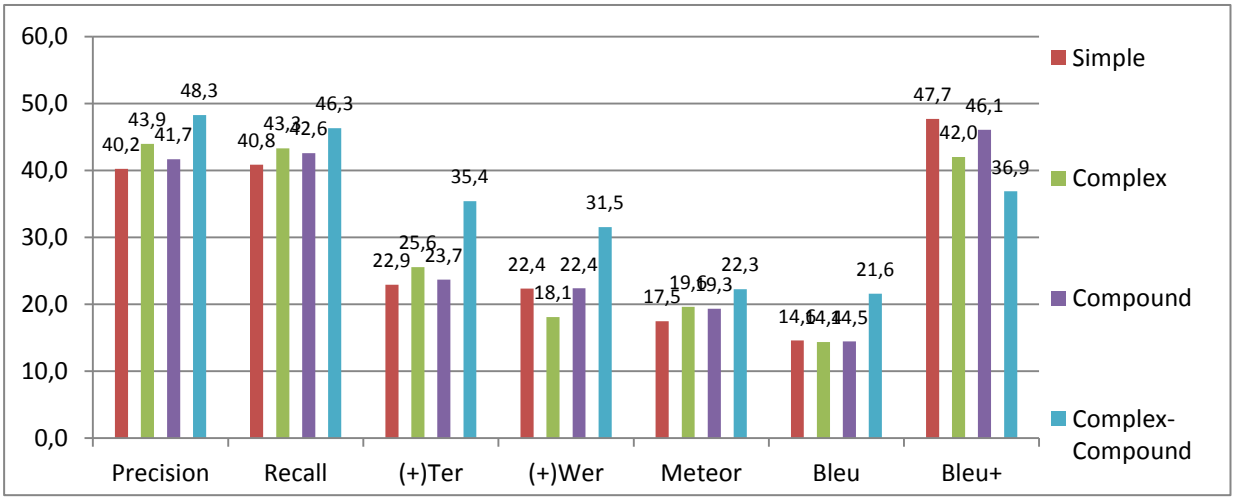


Figure 5.1.5: Average Scores of Evaluation Rates for All Structures of Turkish Sentences on Yandex

It is understood from the figure 5.1.5 that there are many word with same root but different suffix. So Yandex service can catch word root successfully.

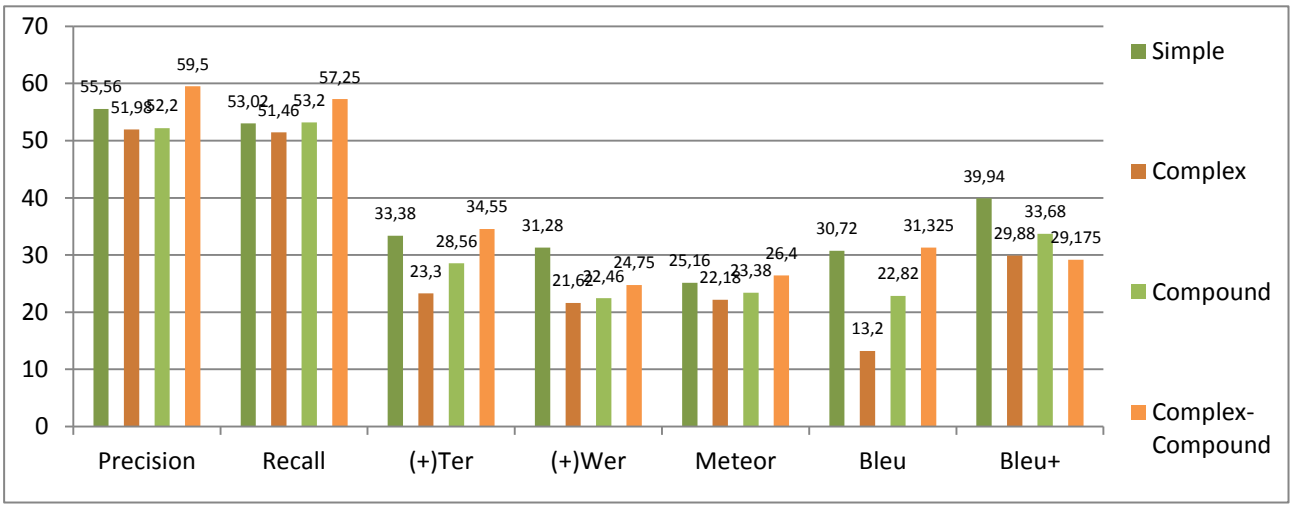


Figure 5.1.6: Average Scores of Evaluation Rates for All Structures of English Sentences on Yandex

It is seen obviously that compound-complex sentences can be translated y Yandex better than other type of sentences from Turkish to English language.

Metric evaluation vs. Sentence structure was evaluated over 3 different bilingual dataset: Google, Bing and Yandex.

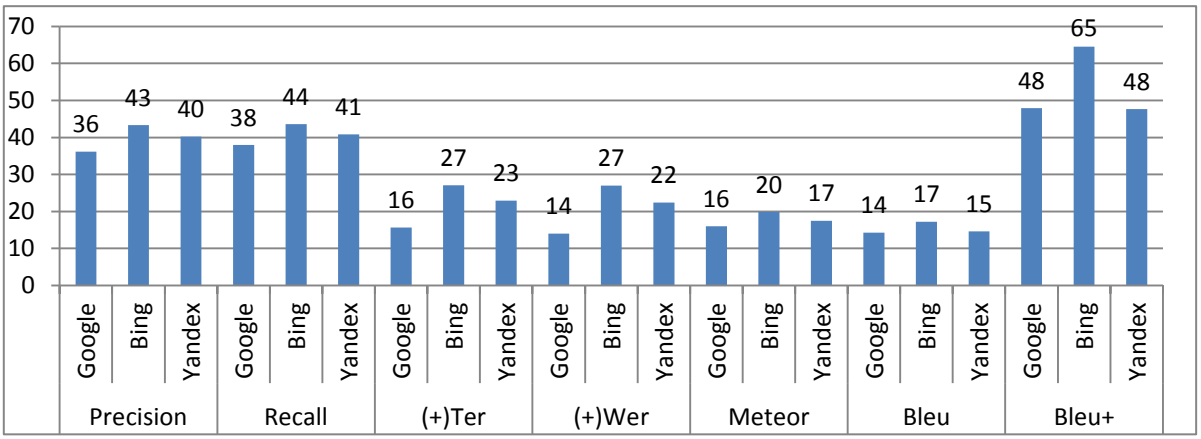


Figure 5.1.7: Average Scores of Evaluation Rates for Simple Turkish Corpus on Services

Comparing all metrics, it is clearly seen that there is big difference between Bleu and Bleu+ rates since word suffixs and synoym usage.

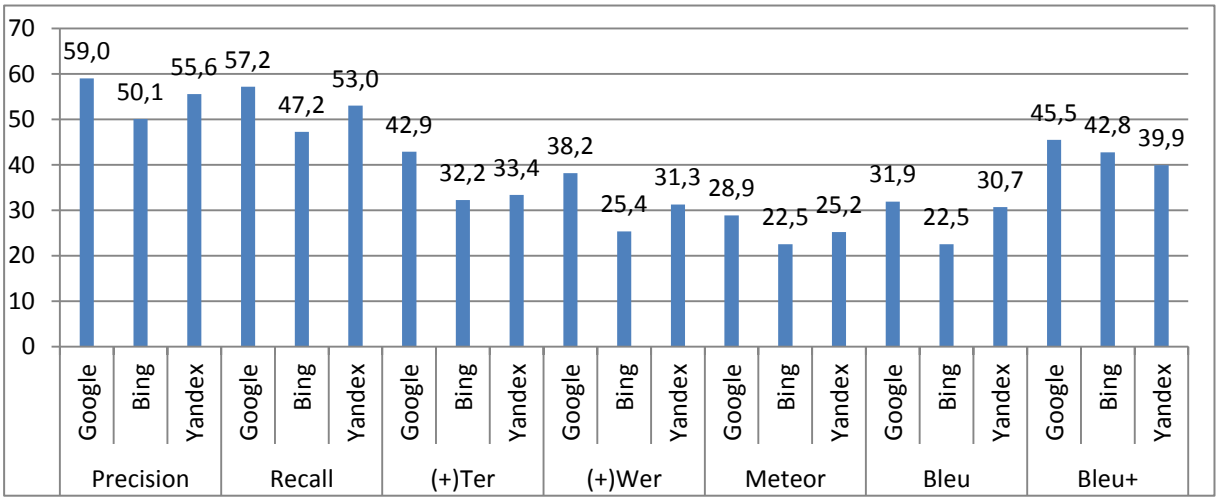


Figure 5.1.8: Average Scores of Evaluation Rates for Simple English Corpus on Services

Over simple sentence from Turkish to English similarity evaluation test, it is seen clearly that Yandex can perform better than Bing.

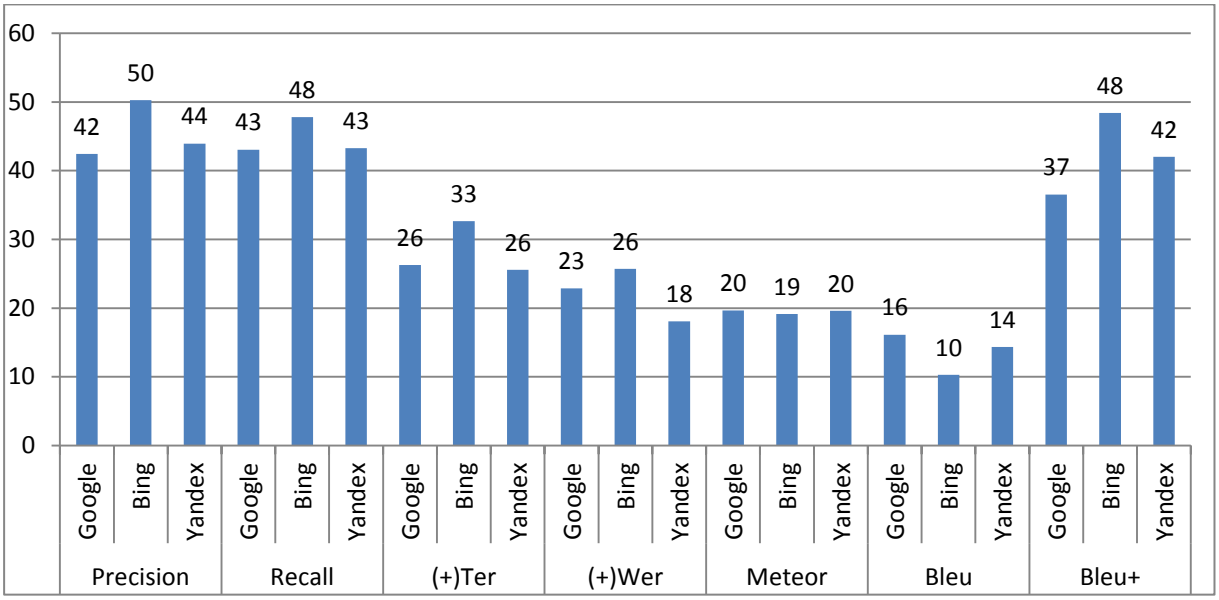


Figure 5.1.9: Average Scores of Evaluation Rates for Complex Turkish Corpus on Services

It can be understood by the Figure 5.1.9 that Bing is much successfully than Google and Yandex on translation from English to Turkish with complex sentences.

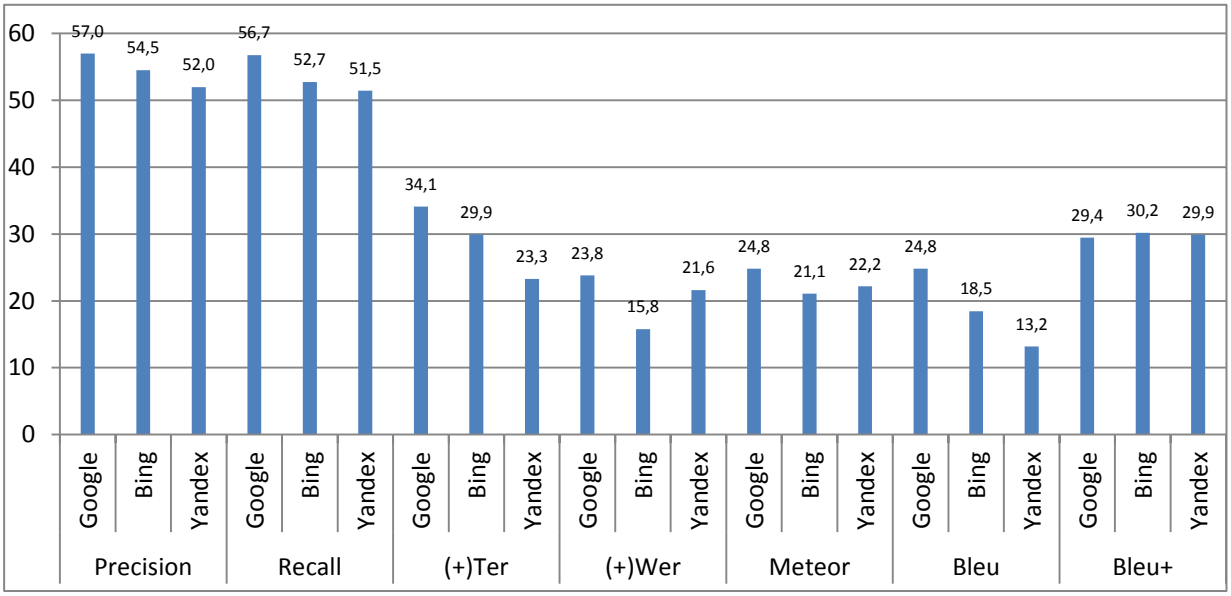


Figure 5.1.10: Average Scores of Evaluation Rates for Complex English Corpus on Services

From English to Turkish almost all services show same quality but a little bit differences. So Google provides better score than Bing and Yandex with a small difference.

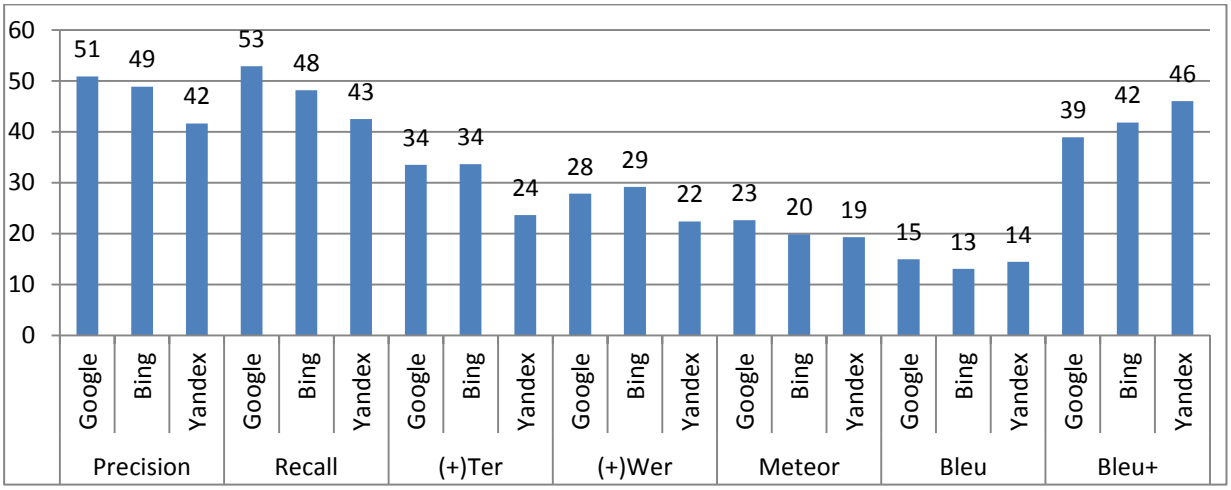


Figure 5.1.11: Average Scores of Evaluation Rates for Compound Turkish Corpus on Services

In viewing metrics comparatively, especially Precision, Recall and Bleu, for complex sentence structure Google gives higher rate than Bing and Yandex.

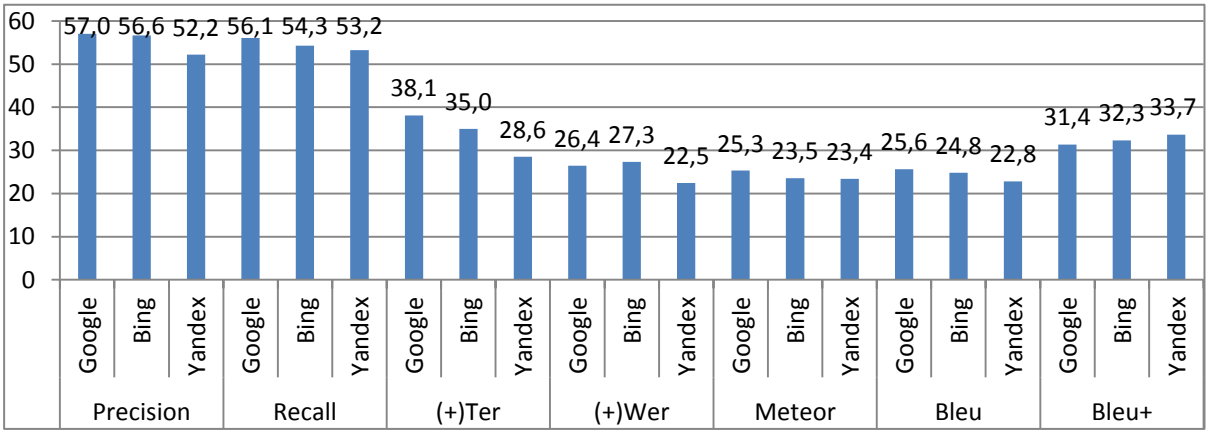


Figure 5.1.12: Average Scores of Evaluation Rates for Compound English Corpus on Services

It is clearly seen that for compound sentences, online translation service exhibit almost same features.

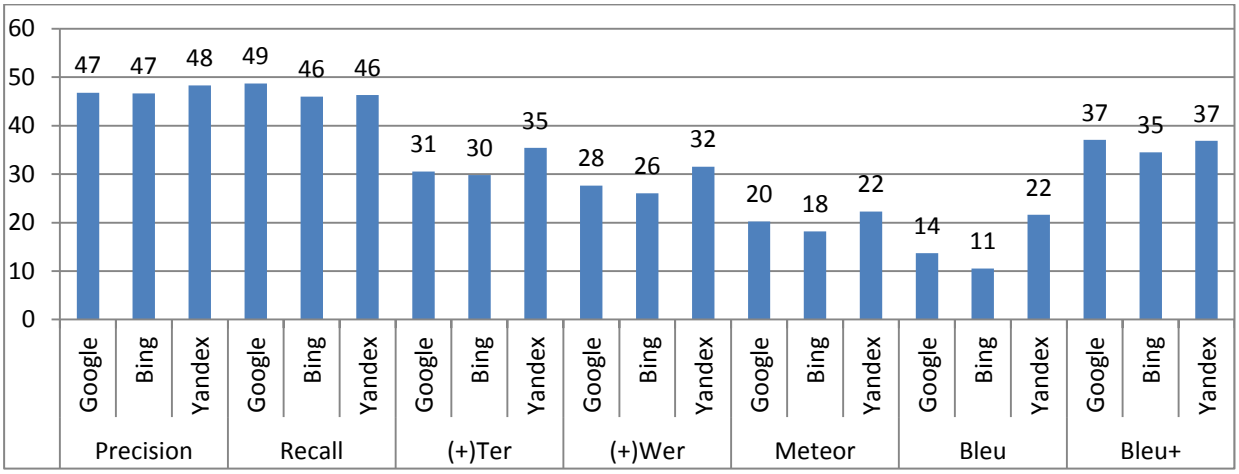


Figure 5.1.13: Average Scores of Evaluation Rates for Compound-Complex Turkish Corpus on Services

It is seen clearly that we can say from English to Turkish translation by Google is giving more same root with different suffixes according to Bleu-Bleu+ comparison.

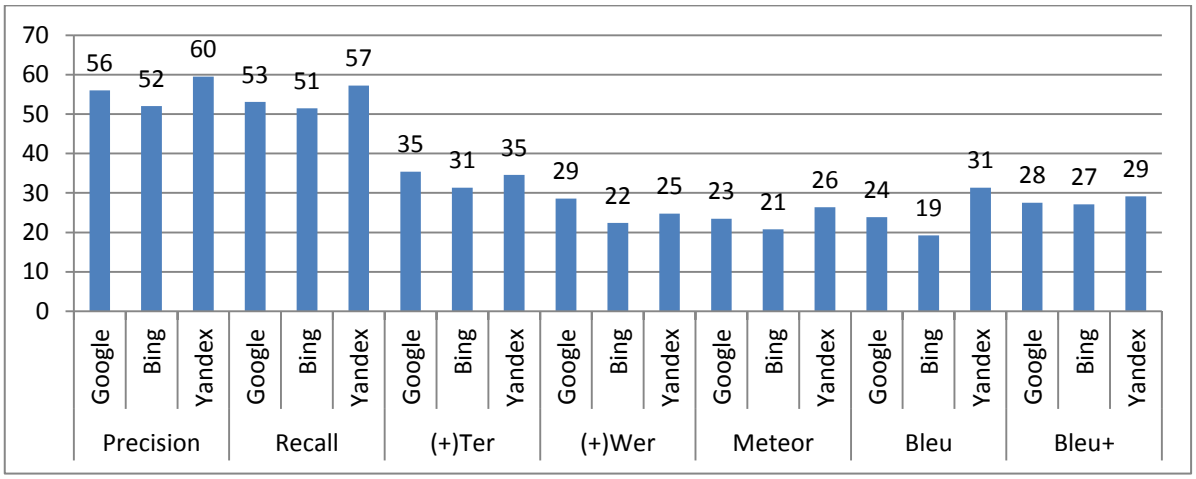


Figure 5.1.14: Average Scores of Evaluation Rates for Compound-Complex English Corpus on Services

For complex-compound sentences evaluation, this figure expose that three online services provide almost same results in terms of every metric with slightly small differences.

5.2 Verification Test Results

The following figures are demonstrating overlap rates coming from verification test result over training test result.

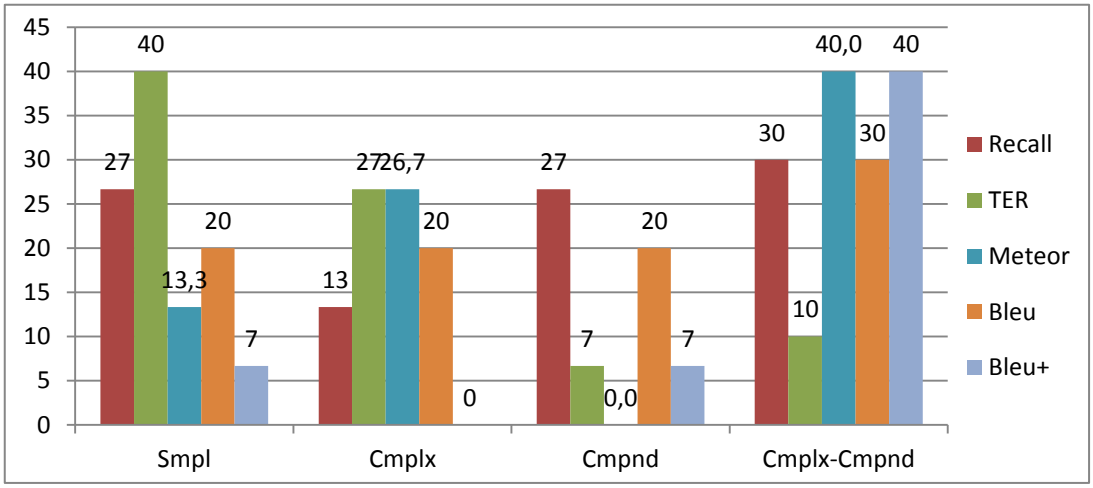


Figure 5.2.1: Evaluation Score of Automatic Verification Test Results for Turkish Corpus on Google

The figure on the above demonstrates that excluding over Meteor metric on compound sentence structure, almost all metric confidence interval are coherent.

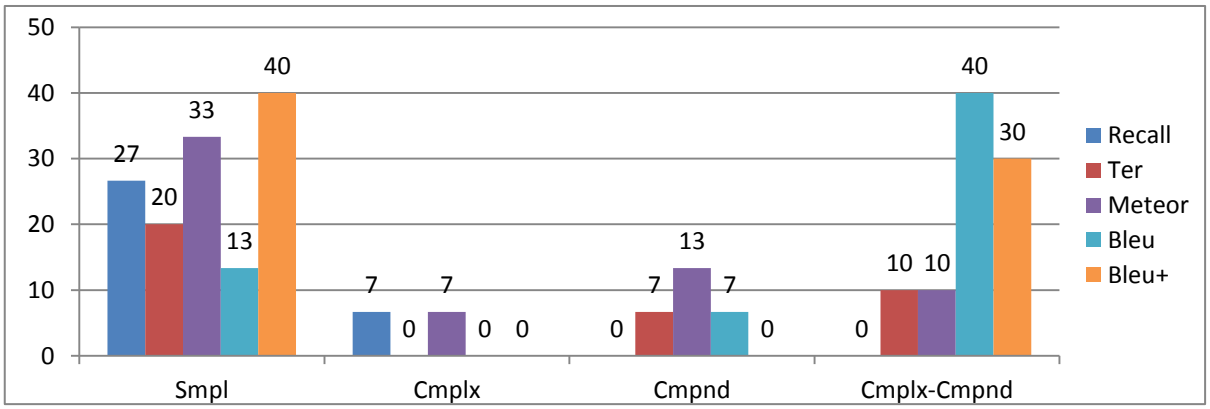


Figure 5.2.2: Evaluation Score of Automatic Verification Test Results for English Corpus on Google

It is seen clearly that from Turkish to English translation verification tests cannot give exactly confidence interval especially complex and compound sentence structure because of their flexible content.

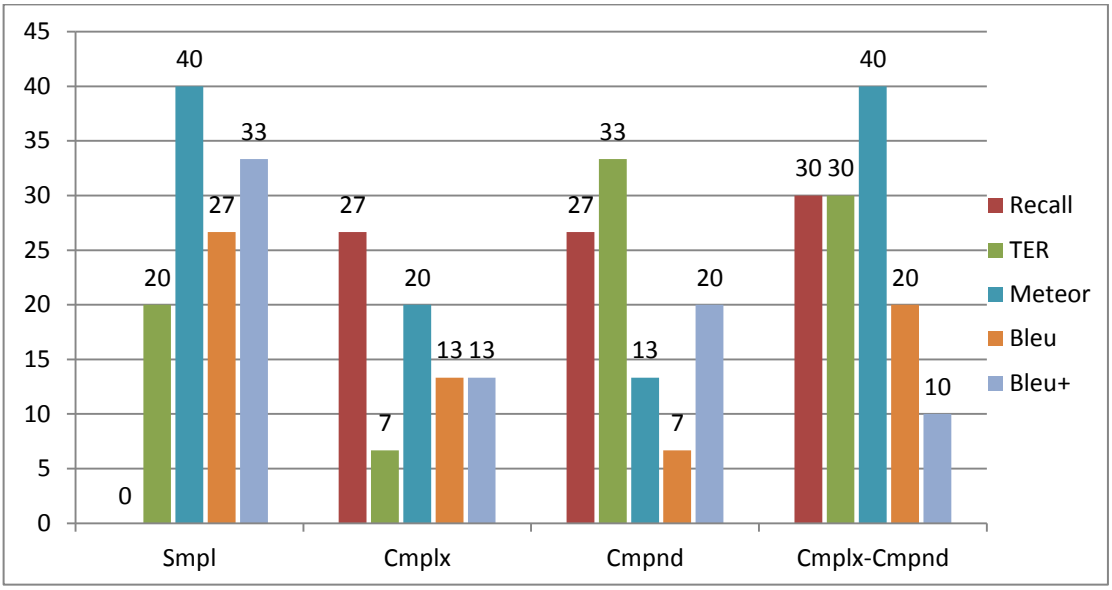


Figure 5.2.3: Evaluation Score of Automatic Verification Test Results for Turkish Corpus on Bing

Over English to Turkish translation, verification of already trained confidence interval gives consistency respectively.

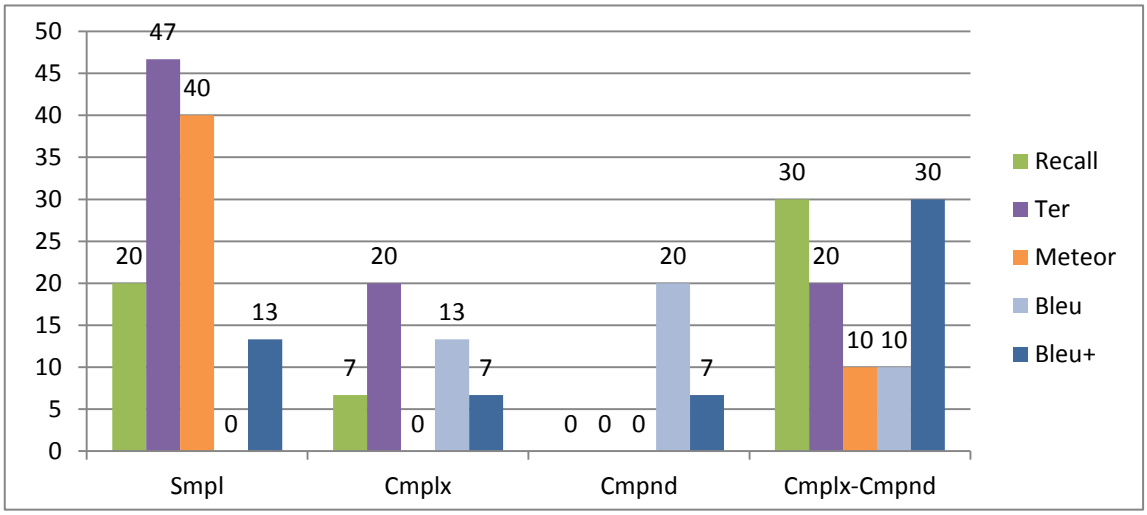


Figure 5.2.4: Evaluation Score of Automatic Verification Test Results for English Corpus on Bing

The most remarkable result in Figure 5.2.4 is that the estimation of confidence interval for both word occurrence and alignment cannot be done together exactly for Bing service over from Turkish to English translation.

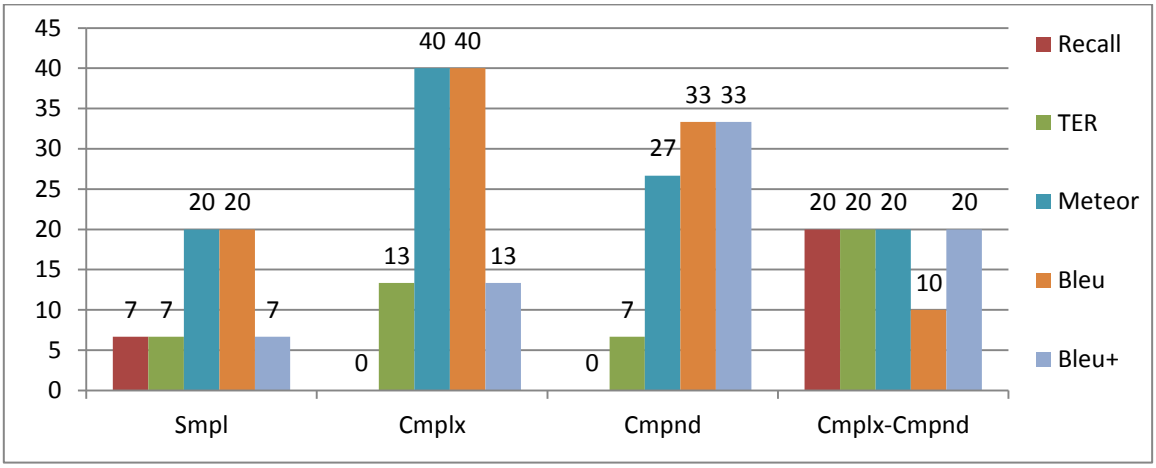


Figure 5.2.5: Evaluation Score of Automatic Verification Test Results for Turkish Corpus on Yandex

Just confidence interval verification test gives us the fact that verification set overlaps of confidence interval cannot be reflect real range of average especially occurrence of word in sentence by recall in from English to Turkish translation. Since, it can be said that estimation of confidence interval depends on corpus attributes.

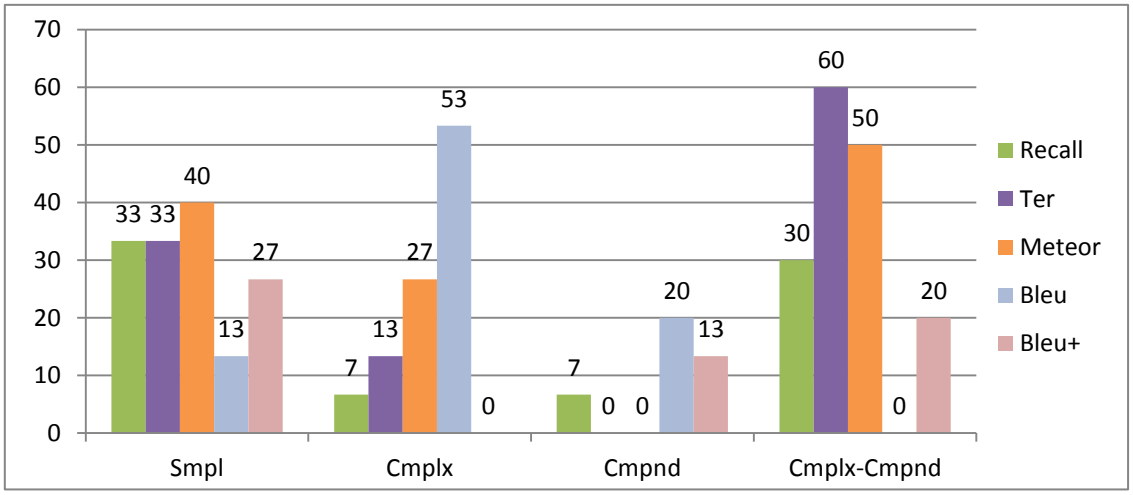


Figure 5.2.6: Evaluation Score of Automatic Verification Test Results for English Corpus on Yandex

The above figures show that especially sentences evaluated with Meteor method give similar results. However, simple and complex-compound sentences give average similarity score. On the other hand complex and especially compound sentences in English languages can be very different with each other.

CHAPTER 6

CONCLUSION AND FUTURE WORK

5.1 Conclusion

In recent years, the need of automatic translation is rising. In addition, people desire to know accuracy rate of translation. So evaluation of machine translation is highly crucial. Wrong translation samples cause to lost prestige. MT is hard; evaluation of MT is even harder.

The aim of this research is to describe machine evaluation techniques and quality results of different online machine translation services. An overview of the current machine translation technology and automated evaluation metrics is given. In this research, especially sentence structures based online machine translate services evaluation such as Google, Bing and Yandex has been used to estimate translation quality levels for Turkish and English languages.

Since there is no bilingual corpus about sentence structures, we decided to create a corpus classified in terms of sentence structure from bilingual sentence collection from academic, historical and news domains. With this special corpus which contains sentences both in English and Turkish languages, each sentence is categorized based on its sentence structure by hand under experts' opinions. The sentences categorized as Simple, Complex, Compound and Complex-Compound.

Moreover, each sentences' automatic candidate translations were obtained from selected online translation services; Google Translate, Bing Translate and Yandex Translate. After data preparation step, this corpus was separated in to 3 different subsets to examine online translation services. First one is used to take a generally estimation average and interval rates.

Second one was used for to try to validate first test results, and last one is used for manual validation test by comparing with human evaluation results.

The corpus categorized based on sentence structure was evaluated by using with some popular similarity metric such as Precision, Recall, Meteor, Bleu, Bleu+, etc. Translation evaluation tests showed us that both machine translation and its evaluation must be special for source and target languages because of languages have different characteristic. So, while one metric is suitable to measure the quality of a language, and has high correlation with human judgment for English language, it might not give closer scores to human evaluation in Turkish translation output evaluation.

Results showed clearly that online translation services perform different quality level over sentence structure on different language and similarity measurement metrics cannot be used for all. Metrics must be selected eligible and suitable for sentence similarity measurement and language specific. And some of them must be used together to be more meaningful.

It is seen clearly that in linguistic scientific field, there is no exactly stable and reliable of confidence interval since various corpus content. But comparatively test showed that the evaluation results of online translation services' output quality gives significant feedback to MT users should not to use them directly by trusting without any paraphrase. Online translation services can be used as a translation helper or pre-translator. So candidate translation coming from machine translators must be post edited by manually to obtain a good correct and easy readable translation.

TER scores are all the time is bigger than WER scores since shift process decrease error points. And negative scores reflect big difference about number of words between reference and candidate text. TER metric always gives bigger score than WER because it is using shift concept. Because of insertion or/and deletion measurement process TER and WER metrics may give scores under zero (negative). These results give evidence about difference number of Word both in reference and candidate sentences. Actually, WER is used in signal processing. So it is not useful and it cannot accommodate to measure outputs quality of translation tools. |Bleu+-Bleu| distance is bigger in Turkish sentences evaluation than English sentences because there are many different suffixes in Turkish language than in English.

Precision is similar to Recall and TER is similar and gives more efficient score than WER according to human approach. So in some figures Precision and WER metric results are not included.

Machine Translation is a hard problem. There are many weaknesses and short comings of machine translators. Translation quality depends on language family, Text category such as structure, type, word length (short, normal, long text), and selection priority of word and idiom meaning, etc. because there are a lot of meanings of words. The state of the art cannot reach on to 100% correctness. Perfection is so far away. But this study showed that online machine translators don't give same quality of candidate translation in terms of language and sentence structure. Results can vary depending on sentences selected.

The free translation services seem to do a better job of handling simple sentences, and some of them appear to be making a serious effort to deal with complex sentences and context rather than translating a word at a time. They are best used when translating from a foreign language into your own, as when you are trying to understand a foreign language website. They should not be used if you are writing in a foreign language for publication.

The quality of translation is dependent on the language pair. Typical errors in machine translations are: missing words, word order, incorrect words, unknown words, and punctuation. Although automated evaluation is fast, cheap, required minimal human labor, and no need for bilingual speakers, current metrics are still relatively crude and individual sentence scores are often not very reliable.

5.2 Future Work

As the technology is getting advanced, the quality of language translation services will get higher in the future. Currently free language translation services are useful for simple sentence structures, but for complex sentence structures human post-editing needed for getting a comprehensive and quality translation of any language.

With the larger and comprehensible corpus, the more real and better correlated with human approach results can be obtained. There may be better evaluation to correlate with human judgment by examining on word prefixes, suffixes, root synonyms, parallel meanings and hybrid metric usage. Automatic sentence structure determiner makes easier to classify

sentences. Other languages and many evaluation metrics can be applied to examine comparing texts. Additionally, using multi reference, true chunking, and part-of-speech tagging and true word sense detection may be so beneficial.

REFERENCES

[1] C. Kit and T. M. Wong. "Comparative evaluation of online machine translation systems with legal texts." *Law Libr. J.* 100 (2008): 299.

[2] C. C. Burch, “Machine Translation”, 22.02.2015, [Online]. Accessible: http://cis.upenn.edu/~ccb/

[3] P. Brown, "A statistical approach to language translation. “Proceedings of the 12th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1988.

[4] Data Sources for and Evaluation of Machine Translation, Trevor Cohn, 31.12.2014, [Online]. Accessible: http://staffwww.dcs.shef.ac.uk/people/T.Cohn/mt/day5-evaluation.pdf

[5] S. Tripathi and J. K. Sarkhel. "Approaches to machine translation." Annals of library and information studies 57 (2010): 388-393.

[6] “Language”, 22.02.2015, [Online]. Accessible: http://en.wikipedia.org/wiki/Language

[7] D. O’Neil, “What is Language? “, 22.02.2015, [Online]. Accessible: http://anthro.palomar.edu/language/language_2.htm

[8] Santa Fe Institute, “Evolution of Human Languages”, 22.02.2015, [Online]. Accessible: http://ehl.santafe.edu/intro1.htm

[9] J. Crowgey “Fundamentals of Grammar: Lexical and Grammatical Categories”, 22.06.2012, [Online]. Accessible: http://courses.washington.edu/ling100/lect_slides/04_lex_grm_cat/04_lex_grm_cat.pdf

[10] Kisno, Fundamentals in Linguistics: An Introduction, Halaman Moeka, Jakarta,2012

[11] “Morphemes in English grammar”, 22.02.2015, [Online]. Accessible: https://www.tesol-direct.com/guide-to-english-grammar/morphemes

[12] T. Tinsley, K. Board, Languages for the Future, 30.11.2013, [Online]. Accessible: http://www.britishcouncil.org/sites/britishcouncil.uk2/files/languages-for-the-future.pdf

[13] E. O'Brien, "Sentence Type", 22.02.2015, [Online]. Accessible: http://www.english-grammar-revolution.com/sentence-types.html

[14] "Conversion of a complex sentence into a simple sentence", 22.02.2015 [Online]. Accessible: http://www.englishpractice.com/grammar/conversion-complex-sentence-simple-sentence/

[15] The Writing Center of University of Central Missouri," TYPES OF SENTENCES: SIMPLE, COMPOUND, COMPLEX, and COMPOUND-COMPLEX", 22.02.2015 [Online]. Accessible: https://www.ucmo.edu/ae/writing/documents/TYPESOFSENTENCES.pdf

[16] "NLP", 22.02.2015, [Online]. Accessible: http://www.webopedia.com/TERM/N/NLP.html

[17] R. Lazerowitz, "What is Natural Language Processing?", 14.09.2014, [Online]. Accessible: http://infospace.ischool.syr.edu/2012/05/11/what-is-natural-language-processing/

[18] "Natural language processing", 15.12.2014, [Online]. Accessible: http://en.wikipedia.org/wiki/Natural_language_processing

[19] "Translation", 14.10.2014, [Online]. Accessible: http://www.oxforddictionaries.com/definition/english/translation

[20] J. Hutchins, "Machine translation and human translation: in competition or in complementation." International Journal of Translation 13.1-2 (2001): 5-20.

[21] O. Karami, "The brief view on Google Translate Machine", 22.01.2015, [Online]. Accessible: http://logic.at/lvas/185054/GoogleTranstaeMachineBriefView_OmidKarami.pdf

[22] P. Koehn, "Statistical Significance Tests for Machine Translation Evaluation." EMNLP. 2004.

[23] M. Simard, "Rule-based translation with statistical phrase-based post-editing." (2007).

[24] H. Somers, "Review article: Example-based machine translation." Machine Translation 14.2 (1999): 113-157.

[25] "WHAT ARE THE MAIN TYPES OF MACHINE TRANSLATION?" 11.11.2014, [Online]. Accessible: http://www.machinetranslation.net/quick-guide-to-machine-translation/machine-translation-technologies

[26] "Comparison of machine translation applications, 22.06.2014, [Online]. Accessible: http://en.wikipedia.org/wiki/Comparison_of_machine_translation_applications

[27] “Foreign Transcript Evaluation and Translation”, 15.11.2014, [Online]. Accessible: http://www.alamo.edu/uploadedFiles/St_Philips_College/Library/Files/eval-translation.pdf

[28] S. Bangalore, G. Bordel, and G. Riccardi. "Computing consensus translation from multiple machine translation systems." Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on. IEEE, 2001.Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[29] M. Smets, M. Gamon, J. Pinkham, T. Reutter and Martine Pettenaro, "High quality machine translation using a machine-learned sentence realization component." Proceedings of MT Summit IX. 2003.

[30] K. Chatzitheodoroua, S. Chatzistamatis “An Open Toolkit for Human Machine Translation Evaluation”, 22.08.2014, [Online]. Accessible: https://ufal.mff.cuni.cz/pbml/100/art-chatzitheodorou-chatzistamatis.pdf

[31] M. Snover, N. Madnan, B. J. D. and R. Schwartz, "Fluency, Adequacy, or HTER? Exploring different human judgments with a tunable MT metric." Proceedings of the Fourth Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2009.

[32] S. Bangalore, O. Rambow, and S. Whittaker. 2000. Evaluation Metrics for Generation. In Proceedings of the International Conference on Natural Language Generation (INLG 2000), Mitzpe Ramon, Israel. 1-13.

[33] S. Nießen, "An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research." LREC. 2000.

[34] C.Tillman, S. Vogel, H. Ney, H. Sawaf, and A. Zubiaga. 1997. Accelerated DP-based search for statistical translation. In Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech ’97), 2667–2670. Rhodes, Greece.

[35] M. Snover, "A study of translation edit rate with targeted human annotation." Proceedings of association for machine translation in the Americas. 2006.

[36] K. Papineni, "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[37] C. Y Lin and F. J. Och. "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics." Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.

[38] M. Rajman and T. Hartley, 2001, Automatically predicting MT systems ranking compatible with Fluency, Adequacy and Informativeness scores. In "Proceedings of the 4th ISLE Workshop on MT Evaluation, MT Summit VIII". Santiago de Compostela, September 2001. pages. 29-34.

[39] A.C. Tantuğ, K. Oflazer, İ. D. El-Kahlout, 2008. "BLEU+: A Fine Grained Tool for BLEU Computation", In Proceedings of Language Resources and Evaluation Conference LREC, Morocco.

[40] G. Doddington. "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics." Proceedings of the second international conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc., 2002.

[41] M. Gonzalez and J. Giménez. "An Open Toolkit for Automatic Machine Translation (Meta-) Evaluation." (2014).

[42] S., Marina, N. Japkowicz, and S. Szpakowicz. "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation." AI 2006: Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2006. 1015-1021.

[43] S. Banerjee, and A. Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005.

[44] M. Denkowski, and A. Lavie. "Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems." Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2011.

[45] I. D.Melamed, "Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons." arXiv preprint cmp-lg/9505044 (1995).

[46] I. D.Melamed, R. Green, and J. P. Turian. "Precision and recall of machine translation." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2. Association for Computational Linguistics, 2003.

[47] T. Gornostay, "Machine Translation Evaluation."25.111.2014, [Online]. Accessible: : http://www.ida.liu.se/labs/nlplab/gslt/mt-course/info

[48] D. Cer, C. D. Manning, and D. Jurafsky. "The best lexical metric for phrase-based statistical MT system optimization." Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

[49] C. C. Burch, "Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk." Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009.

[50] "About SDL", 25.11.2014, [Online]. Accessible: http://www.freetranslation.com/about-sdl.htm

[51] E. Shen,"Comparison of online machine translation tools", 28.11.2014, [Online]. Accessible: http://www.tcworld.info/e-magazine/translation-and-localization/article/comparison-of-online-machine-translation-tools

[52] "About World Lingo", 28.11.2014, [Online]. Accessible: http://www.worldlingo.com/en/company/pr/pr20090817_01.html

[53] "Asiya", 29.11.2014, [Online]. Accessible: http://nlp.lsi.upc.edu/asiya/

[54] "Word Error Rate", 01.02.2015, [Online]. Accessible: http://en.wikipedia.org/wiki/Word_error_rate

[55] "A Short Guide to Measuring and Comparing Machine Translation Engines", 01.02.2015, [Online]. Accessible: http://www.asiaonline.net/EN/Resources/Articles/AShortGuideToMeasuringAndComparingMachineTranslationEngines.aspx

[56] Gupta, V., Joshi, N., & Mathur, I. (2013, August). Subjective and objective evaluation of English to Urdu Machine translation. In Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on(pp. 1520-1525). IEEE.

[57] Shi, C., Lin, D., Shimada, M., & Ishida, T. (2012). Two Phase Evaluation for Selecting Machine Translation Services. In LREC (pp. 1771-1778).

[58] "Which Online Translator Is Best?", 05.02.2015, [Online]. Accessible: http://spanish.about.com/od/onlinetranslation/a/online-translation.htm

[59] Tantug, A. C., Oflazer, K., & El-Kahlout, I. D. (2008). BLEU+: a Tool for Fine-Grained BLEU Computation. In LREC.