The introduction of criteria for assessing an aligned parallel Persian-English corpus at the sentence level

Masoumeh Mashayekhi
Computer Engineering Department
Iran University of Science and Technology, Tehran, Iran

Morteza Analoui
Computer Engineering Department
Iran University of Science and Technology, Tehran, Iran

Behrouz Minaei-Bidgoli
Computer Engineering Department
Iran University of Science and Technology, Tehran, Iran

Abstract - Bilingual corpora are a collection of writings that serve as an example of the relationship between two languages for linguistic and translational applications. Examining the effectiveness of the corpora is one of the essential requirements for working with them. Therefore, the validity of the work based on the corpora is to check their quality. Scientists have identified four main attributes for corpses. These four features are representativeness, limited size, machine-readable shape, standard reference. To evaluate an entity, we need to evaluate these four properties. The limited size and intelligibility of a machine in electronic compartments are certain because they are otherwise unusable. Representativeness means to put a sample set of language variations for the language in question in the corpus. In fact, the corpus has a linguistic diversity. To evaluate this property, we examine the complexity and diversity of the figure and compute the degree of compliance with Ziff's law. For the standardization of each pair, we combine several of the following characteristics: alignment, translation, command, punctuation, separation, characterization. Finally, a fuzzy system uses the final evaluation of these criteria and uses a fuzzy rule base and fuzzy inputs of the introduced evaluators to obtain a fuzzy result for the quality of the entity.

Keywords: evaluation, bilingual corpora, aligned corpora, word alignment, machine translation

1. Introduction

A set of texts is assumed to represent a language or subclasses of the language to allow for linguistic analysis. A corpus is a collection of examples of language texts stored in an electronic form and selected based on external criteria to represent as much as possible a language or language transformations and serve as a source for linguistic research.

In modern linguistics, the corpus can be defined as a body of the natural occurrences of language. In addition, computer corpora are a set of text elements that are compiled for a particular purpose. Often, this huge textual collection is gathered to represent a textual language.

One of the main prerequisites for validating works based on the corpora is to check their quality. The corpus analysis process is very similar to the process of making a corpus. Like a corpus constructor, the corpus analyst must consider the factors, such as; whether the analyzed corpus has a suitable length for specific linguistic studies, and whether the samples within the structure are balanced and have the property of representativeness. Nowadays one of the most common ways of reviewing the results of constructing a corpus is to use it in machine translation.

The less-considered issue is to evaluate the quality of the corpus after it is created. Since the process of generating a translation model by an entity, as well as evaluating the resulting translation, is time consuming and costly and requires many system resources, in this project we try to improve the quality of a structure without creating a translation model. Then we assess the quality of the translation generated by the generated translation model. Hence, in this research, we will extract the effective properties of aligned corpora. In this way, we can provide a method for evaluating the efficiency of the structures that do not require the production of a translation model from the body.

In this research, the basic features of corpora are identified and become accessible to the criteria. For each criterion, a method for quantifying the criteria has been introduced. We have made a few examples of these criteria using separate software. Then a software was created to collect the data set based on the introduced features. Using the created dataset, a fuzzy rule base was designed. Then, using this rule base and fuzzy inference, a general evaluation was made for the corpora.

In the present research, the first part introduces an introduction to the subject of the research; in the second part, the previous methods presented for the evaluation of the entities are mentioned. Part three describes the evaluation method presented for bilinguals. Section four illustrates the results obtained from the implementation of the software, and, the summary and future works are dealt with in Section five.

2. The review of the related literature

The method of using corpus precedes Chomsky's time. First, they used the body for linguistic studies and linguistic structures. According to available reports, the corpus linguistics was much expanded in the early twentieth century [1].

Although today's computer technology has made it possible to provide larger corpora than Chomsky's time, his crisis of the possibility of corpus deviation is an important point to be taken seriously. To resolve this crisis, attempts have been made to choose selective texts in the corpus to have the property of representing linguistic diversity [2]. No further cultural projects can now be found that do not utilize linguistic structures and libraries. [3]

The creation of language databases is another aspect of the use of linguistic structures, with numerous examples of which are now continuously distributed throughout the world. For Persian language, such a base has been created at the Humanities Research Institute [4].

Language monitoring programs also benefit from linguistic features in order to track linguistic developments. Such corpora are called dynamic corpus or monitor corpus [5]. An analytical framework can be created for evaluating linguistic structures in several ways. In the empirical method, a collection of text attributes is selected being agreed upon by its users. This can be done to increase reference efficiency or for other reasons [6]. One way to evaluate the quality of the corpora is to examine the results of their use in the application, for example, we can use them in a translation machine to evaluate bilinguals, and then we can evaluate the results of the translation using the corpus. [7 and 8].

In many cases, they use precision and refinement to evaluate an aligned unit. In these projects, a corpus is selected as the golden standard, and according to this standard, the accuracy, and readability of the aligned texts are calculated [9].

In this project, some of these areas are applied to the English-Persian structure in order to examine the quality of the corpus in some categories. Due to differences between Persian and English, some of these characteristics will be subject to changes. Many bilingual entities have been made, one of the two languages used in them is English. Such as English-Chinese [10], English-French [11], English-Hungarian [12], Swedish-English [13], and many other languages.

3. Research method

The corpus to be evaluated in this project is an aligned English-Persian corpus at the level of the sentence. This corpus is aligned in a volume of one million sentences. The sentences in the corpus are aligned semi-automatically. Several classic literature books and their translations have been used to extract sentences. The sentences are derived from books such as Anna Karenina, David Copperfield, Don Quixote, and their translations.

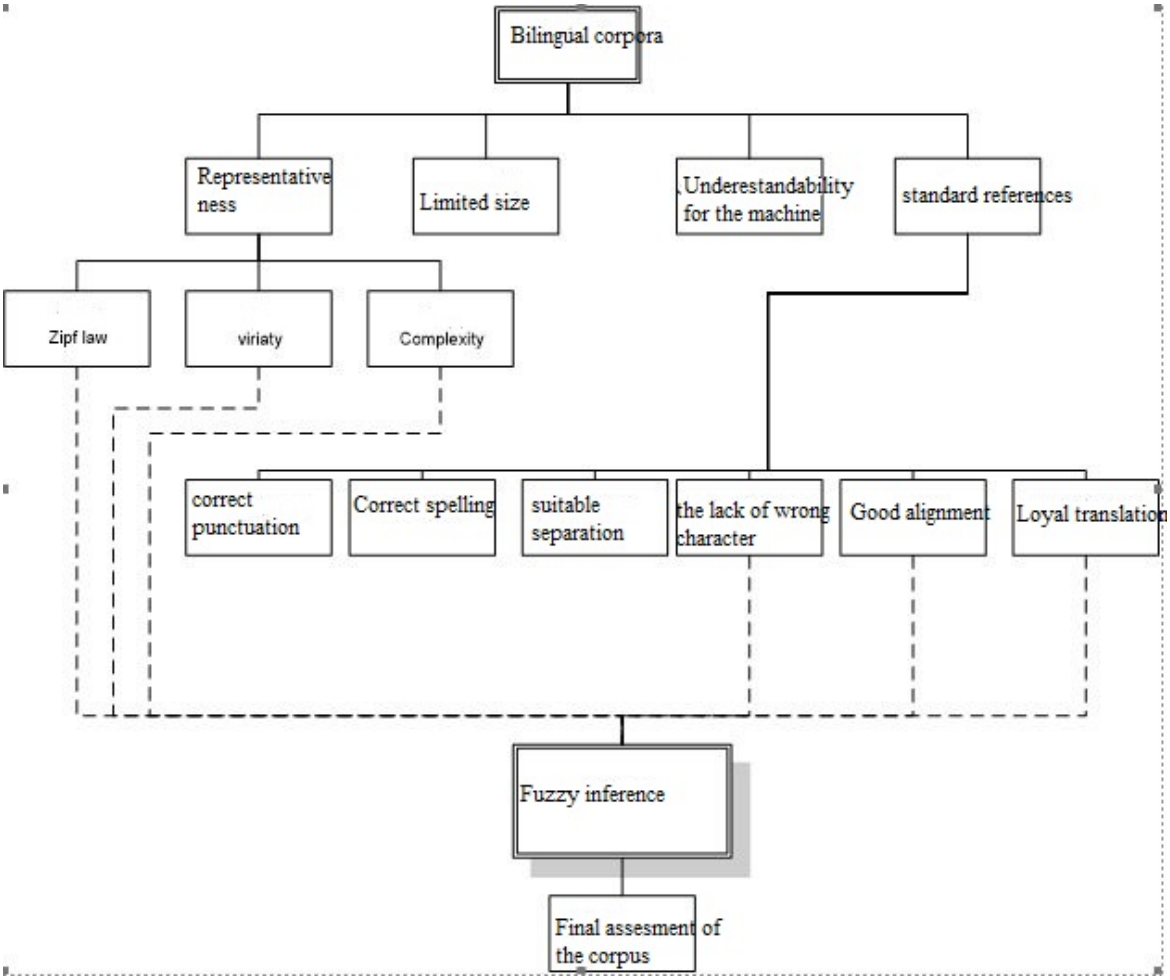


Figure 1: The relationship between selected features and corpus quality

The evaluation of the corpus in this project is limited size and stored in an XML format so that it is understandable to the machine. To evaluate the structure, four main features of the corpus, being mentioned in references and books as representatives of the corpus, are examined. These four characteristics are representativeness, reference standard, understandable for the machine and limited size.

The last two qualities for bilingual corpora are definitely there. Therefore, we examined two features of standardization and representativeness of the corpus. For each of these properties, we select and review the features. For being standard, six linguistic features are selected; good alignment, lack of wrong characters, proper separation, correct spelling and correct punctuation. For being representative, the complexity of the body, its diversity, and the correctness of the repetition of words in the corpus were evaluated.

Arianpur English to Persian dictionary was used to calculate the faithfulness of the translation. The dictionary contains about 50,300 English words along with their translation.

If we use the data obtained from the above-mentioned attributes as inputs of an inference and a rule base, one can make a final evaluation for the corpus. We first define the language terms for each feature. The language features of the terms are used from the part of the corpus created manually. After defining language terms, we transform the data into a fuzzy language term. Then we create fuzzy rules. Fuzzy output terminals are also determined. Finally, using the Mamdani method for inference and the total center for difuzzify, we arrive at the final result. The final result is a number between zero and one that can be determined by the degree of its membership in the output phase' terms.

If the fuzzy inputs of an R_c rule are non-set, such as $A' = u_0$, $B' = v_0$, the resulting degree αi is equal to the minimum value between $\mu_A(u_0)$ and $\mu_B(u_0)$.

(1)

$$\mu_{C_i}(w) = \alpha_i \wedge \mu_{C_i}(w) \quad for\ R_C$$

$$where\ \alpha_i = \mu_A(u_0) \wedge \mu_B(v_0) = \min[\mu_A(u_0), \mu_B(v_0)]$$

In the mamdani method, $\alpha_i \wedge \mu_{C_i}(w)$ is defined as the minimum value.

A class fuzzy inference has been designed that shows the diagram of this class in figure (2). The mfc class has been created to hold each fuzzy term that defines the first and the last values of the first language term, the value of which is one degree of membership, and the language term is defined in its specification. Functions for this class are used as written values in order to get the right and left values for a membership degree (the right and left values are two values that have a membership degree in the fuzzy term) and obtain the membership grade. The conseq class is to store the result of a fuzzy rule. Because the result of a fuzzy rule is a trapezoid, the characteristics for storing this trapezoid are designed.

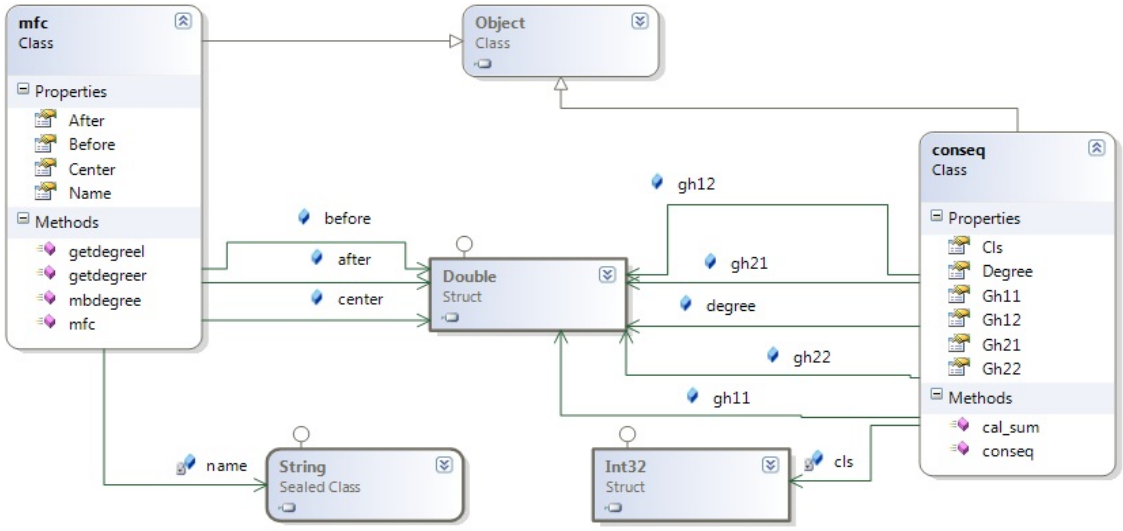


Figure 2: Diagram Class for Fuzzy Inference

Other functions were also used to create inference. In Figure 3, you see the function written for creating a rule. This function takes the input terms of a rule with a literal term of the result of the rule as well as the numerical values of the input of the inference as the input argument displaying an object of the conseq class as output. If the grade in the conseq class is non-zero, then this rule is clear.

```
private conseq rule_resoning(string ccTrans, string ccAlign, string ccChar, string cccompl, string ccvarie, string cczipf,
{
    conseq q;
    int k;
    for (k = 0; k < 5; k++)
        if (mbf_res[k].Name == ccResult)
            break;
    double d1, d2,d3,d4,d5,d6;
    d1 = fuzzify(Trans, mbf_Tras, ccTrans);
    d2 = fuzzify(Align, mbf_Align, ccAlign);
    d3 = fuzzify(Char, mbf_Char, ccChar);
    d4 = fuzzify(compl, mbf_Compl, cccompl);
    d5 = fuzzify(varie, mbf_varie, ccvarie);
    d6 = fuzzify(zipf, mbf_Zipf, cczipf);
    double min = FindMin(d1, d2, d3, d4, d5, d6);
    q = new conseq(k, min, mbf_res[k].Before, mbf_res[k].After, mbf_res[k].getdegreel(min), mbf_res[k].getdegreer(min));
    return q;
}
```

Figure 3: Inference function of a rule

Finally, we give all the results of the rules that are turned on to the non-fuzzy function.

4. Results

4.1. Testing the method of aligning words

To test the implemented method for aligning words using the dictionary, this alignment was done for 6854 couples in a sub-corpus.

The English sentence and the Persian sentence are a pair of corpus in the following.

I remember the precise moment, crouching behind a crumbling mud wall, peeking into the alley near the frozen creek.

دقیقا آن لحظه یادم مانده؛ پشت چینه مخروب‌های دولا شده بودم و کوچه کنار نهر یخزده را دید می‌زدم.

Table 1: aligned results obtained from the software

remember	یاد ماند	peek	دید زد
Precise	دقیقا	alley	کوچه
moment	لحظه	near	کنار
Crouch	دولا شد	freeze	یخزده
Behind	پشت	creek	نهر
crumble	مخروب	i	م
mud wall	چینه		

The only word in this alignment having no equivalent is the word 'into', so the percentage of words aligned is 94.7368.

Another example of a corpus sentence that has certain words is written in the software interface being shown in Figure 4.

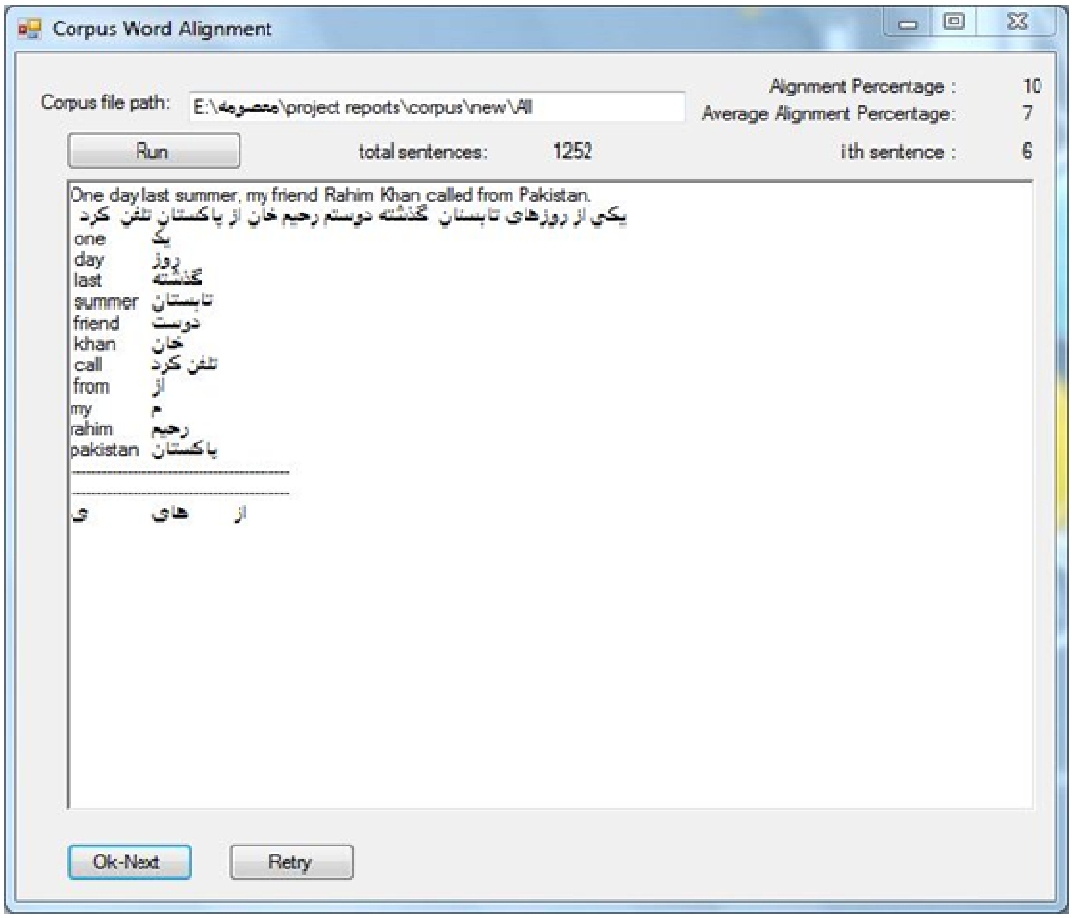


Figure 4: A glimpse of software written to find alignment of words in pair sentences

The table below shows the number of sentences with their alignment percentages.

Table 2: Test data of the method of aligning words using dictionaries

Percentage of words aligned	Number of pairs in corpus	Percentage of corpus pairs
0-10	13	0.18967
10-20	2	0.02918
20-30	13	0.18967
30-40	51	0.744091
40-50	194	2.830464
50-60	712	10.38809
60-70	1433	20.9075
70-80	2225	32.4628
80-90	1568	22.87715
90-100	643	9.381383

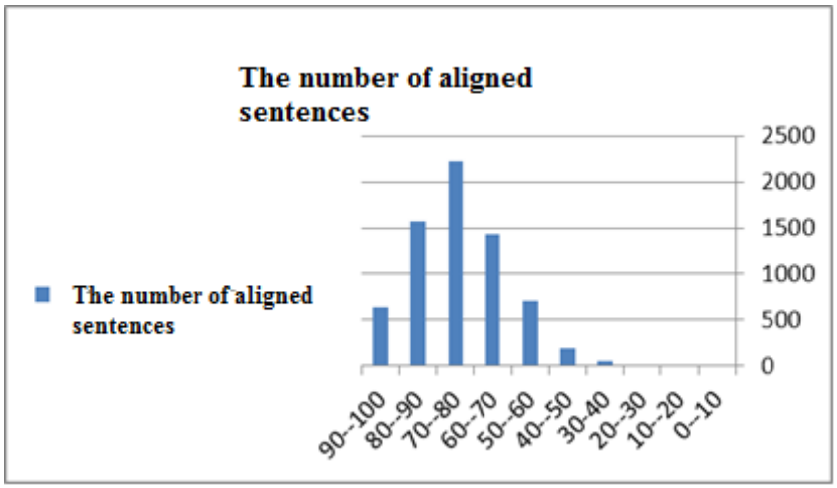


Fig. 5 shows the number of pairs based on the alignment percentage as follows

4.2. Evaluation of the sub corpora and definition of fuzzy terms

In order to have a data set for creating a base of fuzzy rules, a part of the corpus previously evaluated containing 6100 aligned sentences was divided into 100 sentences (about 2000 words for each language). Then, translation evaluation, alignment assessment, character evaluation, evaluation of the complexity of sentences, evaluating sentence diversity, and evaluating the ZIFF law were carried out. The results of the review are as follows.

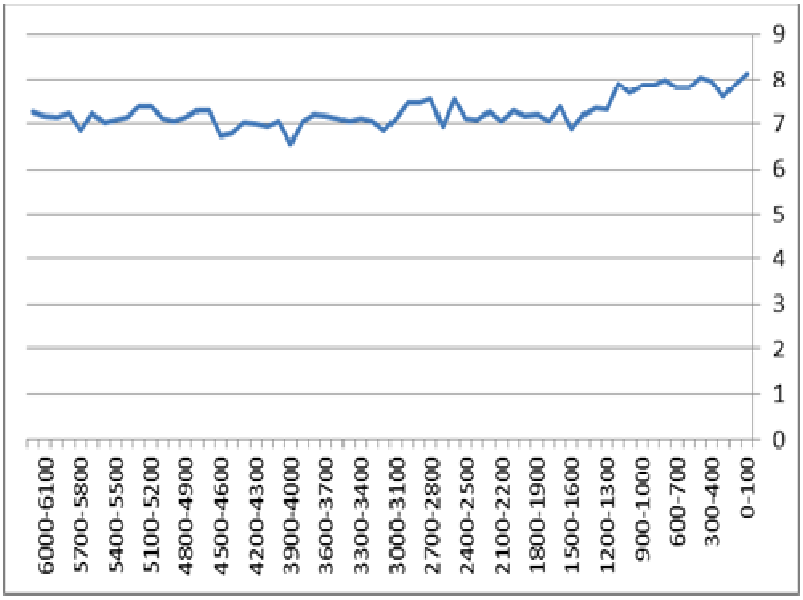


Figure 6: The graph of the percentage of aligned words divided by 10 for the selected section of the figure

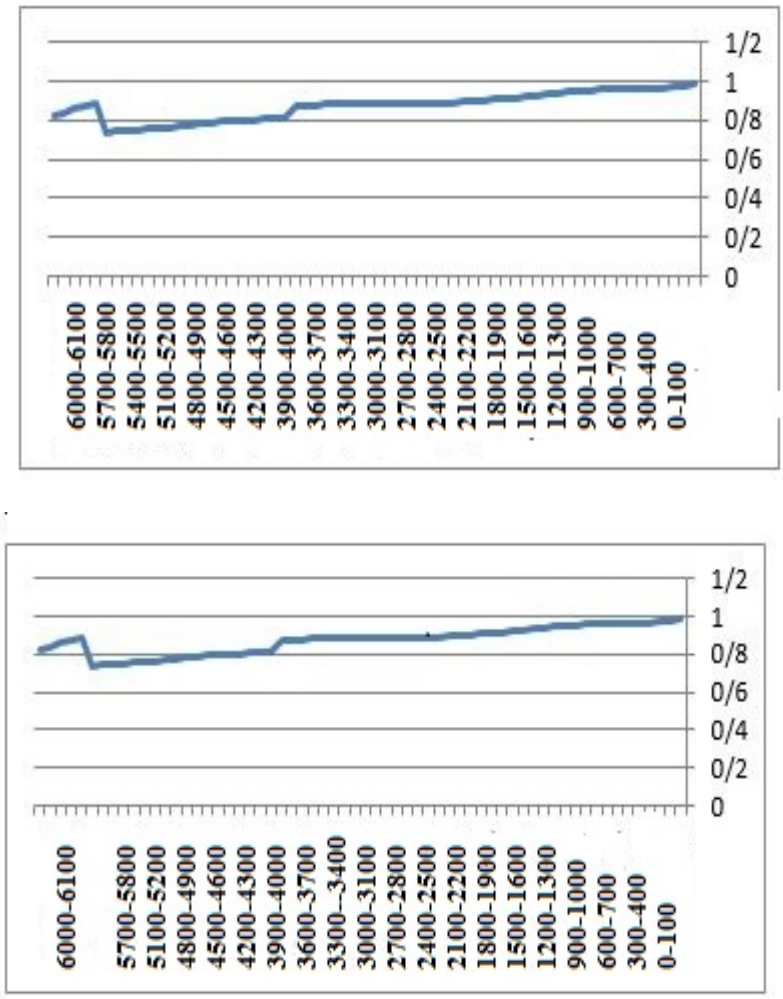


Figure 7: The English-Persian variation ratio for the selected section of the corpus

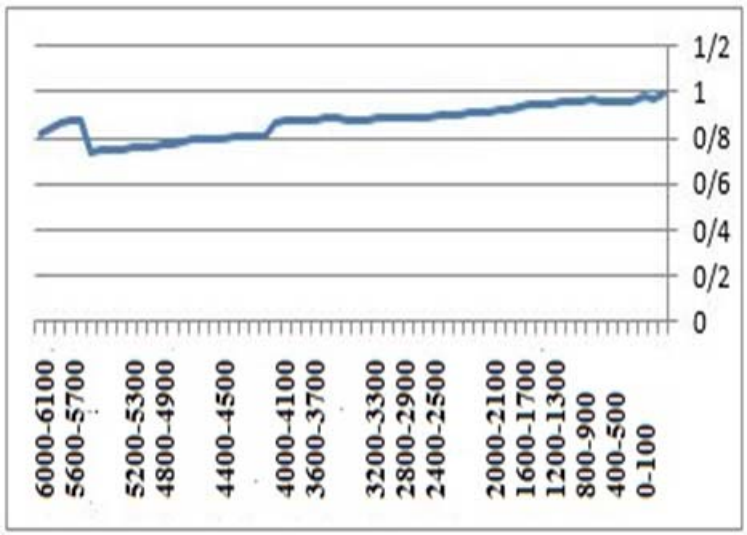


Figure 8: Averages of the wrong characters for the selected section of the corpus

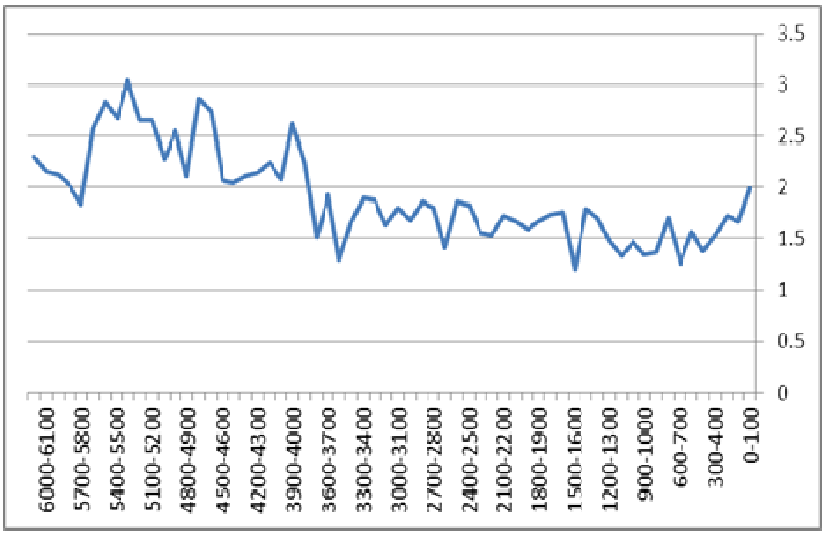


Figure 9: Difference of sections in the marker hypothesis for the selected section of the corpus

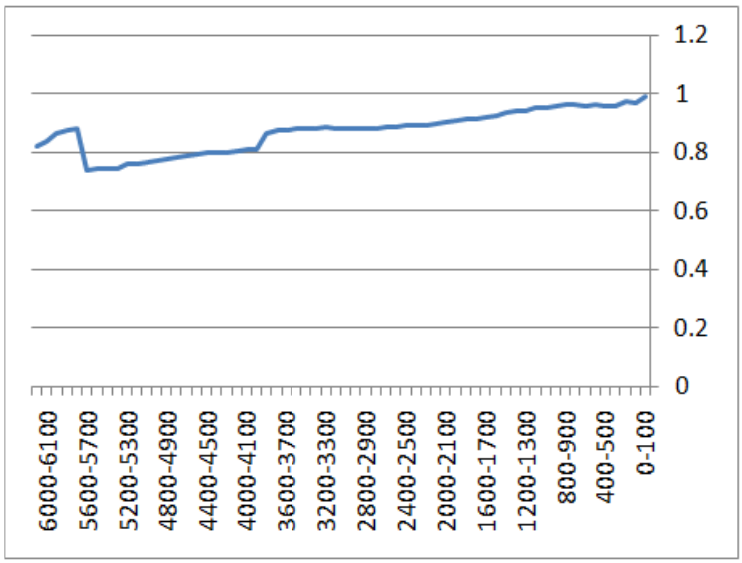


Figure 10: The complexity ratio diagram for Persian and English sentences for the selected section of the corpus

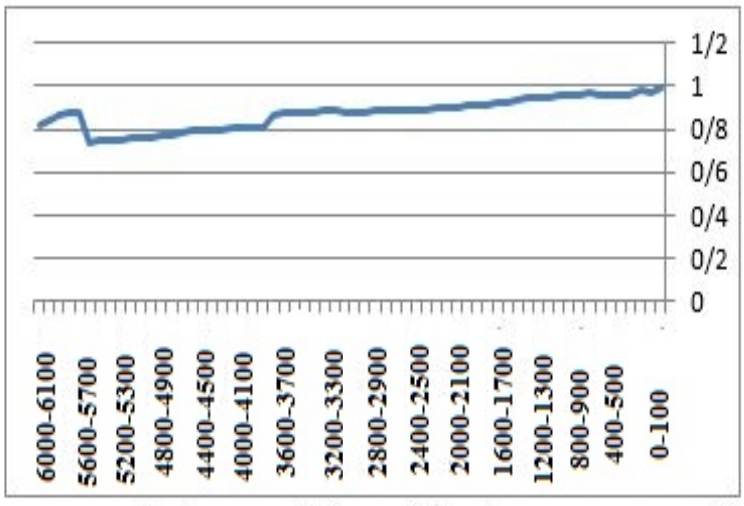


Figure 11: The distance of the English and Persian section of the Ziff rule for the selected section of the corpus

The above graphs identify the right range for each of the data. That is, you can get a "good" fuzzy term for each of the above features. According to the plotted charts, the following values can be defined for each of the "good" fuzzy terms of each property as follows:

1. Average percentage of words aligned in sentences: above 70%
2. The versatility of the sentences in Persian and English: above 0.8
3. Average wrong characters: Under 1 character
4. The average of the difference in the number of sections in the marker hypothesis: under 2 sections
5. Complexity ratio of sentences in Persian and English: above 0.8
6. Distance ratio of Ziff Law: above 0.8

Given the above data, fuzzy linguistic terms are defined. For three characteristics, the ratio of diversity, complexity, and distance from the Ziff law, whose good value is greater than 0.8, is defined as (12). As you can see, fuzzy terms are defined as triangles.

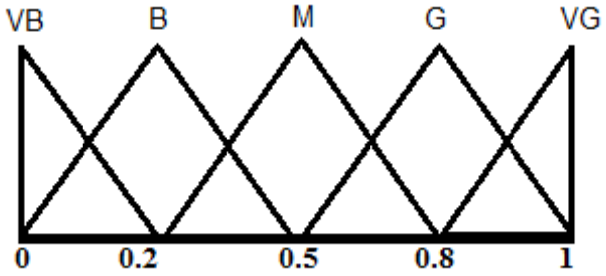


Figure 12: Fuzzy Terminology for Three Characteristics of Complexity, Variation and distance of Ziff Law

In this form, VG means very good, G means good. M represents the middle and VB and B respectively represent very bad and bad. Figure (13) also shows the fuzzy terms corresponding to the percentage of words aligned.

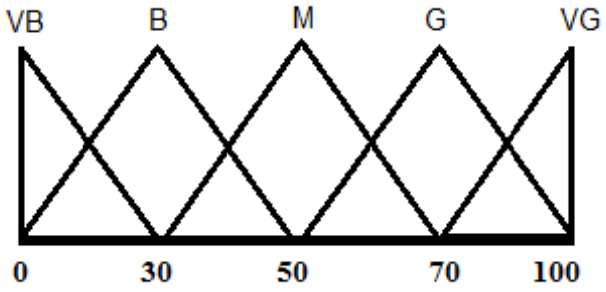


Figure 13: Language terminology to convert the percentage of words aligned to the fuzzy term

For the other two features, we also define fuzzy terms. Therefore, we can convert fuzzy inference inputs to a fuzzy term.

To use deduction, we need to create rules base. Approximately 1,600 bases are used to evaluate the final figure. These rules are based on the policy of the average input language terminology. In other words, if a very good fuzzy term is four, a good fuzzy term is three, medium is two, very bad is one and bad is zero, we can obtain the final output based on the average of the input language terms. For example, if three inputs have a very good result and the other three inputs have a very bad result, then the final output will be fuzzy.

Table 4: An example of the rules used for fuzzy inference

The percentage of words aligned	Marker hypothesis	Average wrong characters	Complexity ratio	Diversity ratio	Distance from Ziff Law	Assessment result
VG	VG	VG	VG	G	G	G
VG	VG	VG	VB	VB	VB	M
G	G	G	G	G	G	G
M	VB	VG	VG	B	B	M
VB	B	G	VG	VB	M	B
M	M	M	B	B	VG	M
B	B	VB	VB	VB	VB	VB
G	G	G	VG	VG	VG	G
VG	VG	VG	VG	VG	VG	VG
VG	VB	B	G	B	B	B

In Table (4-1), VG means very good, G means good, M means medium, and B means bad. In addition, a very bad concept can be considered by VB. For the evaluation result, linguistic terms were determined as Fig. 14.

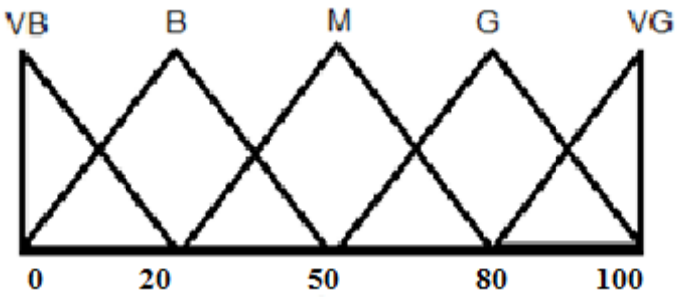


Figure 14: Language terminology for the final evaluation

4.3. Extraction of features from the final shape

The corpus to be evaluated included 67 files in XML format separately. An assessment was made for composite files individually. This file contains 378202 sentences. You can see the distribution of sentences in part of this figure in Fig. 15. The results obtained from applying the methods described on this figure are also mentioned below.

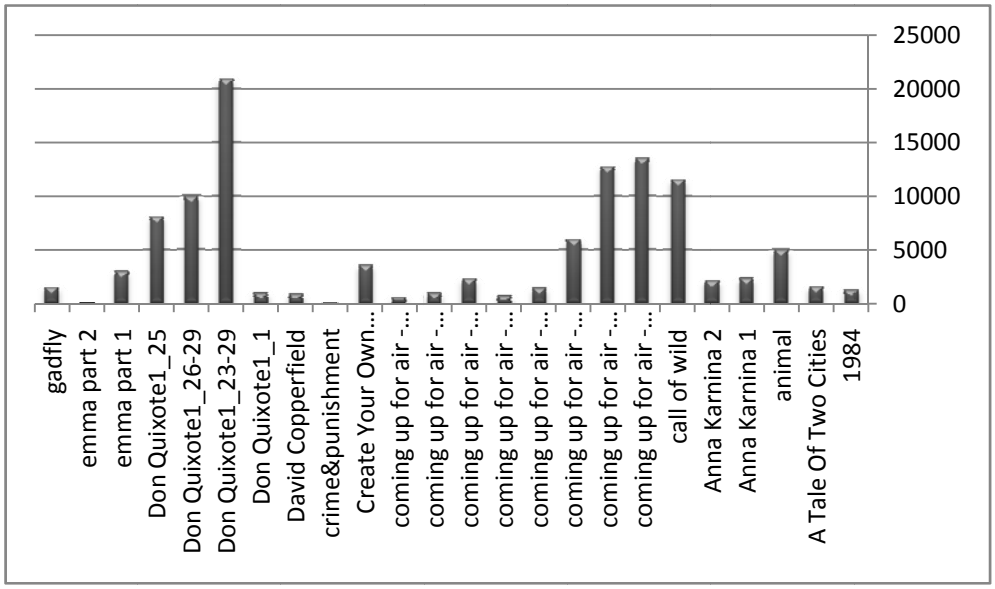
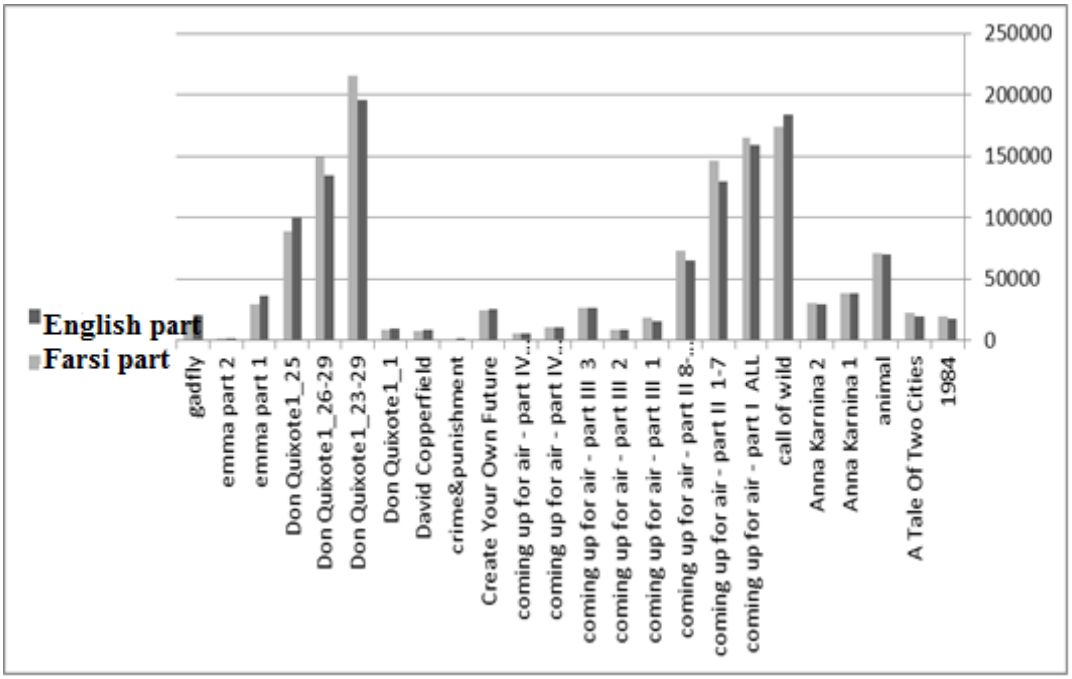


Figure 15: Distribution of sentences in multiple files of the corpus



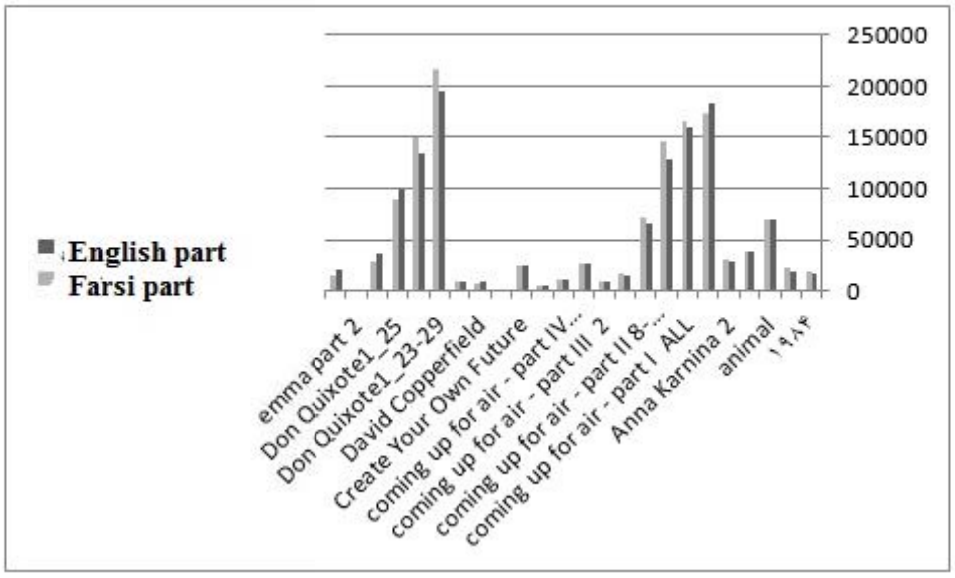


Figure 17: Complexity of Persian and English sections separately

Figure (17) represents the complexity of both Persian and English sections. As is clear in the figure, the complexity ratio in these two parts is numerically close to one. The average complexity ratio in the completely evaluated figure is numerically equal to 937237206.

Figure (18) shows diversity in two parts of English and Farsi for multiple file selections. This graph also indicates a slight variation in the two parts. The average of the diversity of the whole corpus was estimated to be 922682429.

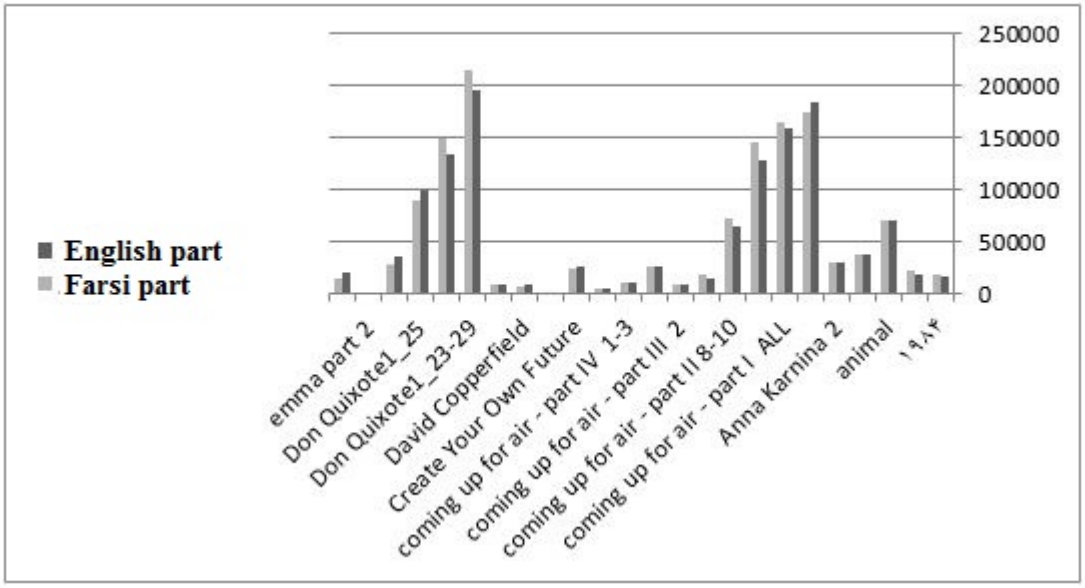


Figure 18: Variety of English and Farsi sections for multiple file formats

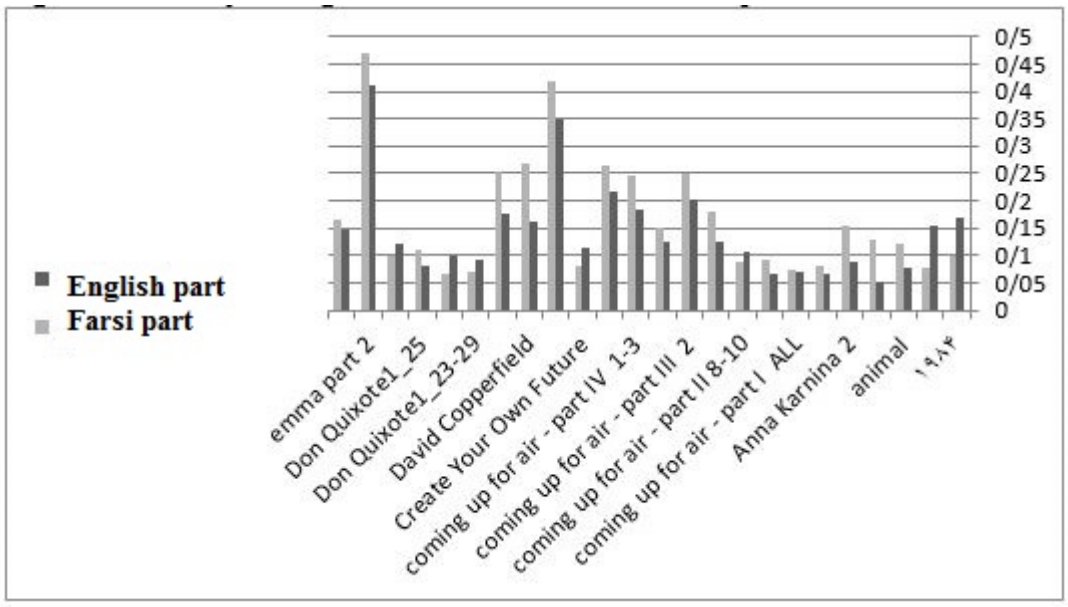


Figure 19: Distance from Ziff Law in Persian and English parts of some files of the corpus

Figure (19) shows that there is a difference between words in the Persian and English sections. Therefore, the average value for the ratio of distance from the Ziff law in English and Persian parts and for the whole corpus is equal to 7449437157.

Figure (20) represents the percentage of words aligned using the dictionary for a part of the figure to be evaluated. For ease of display, the number is divided. The average percentage of words aligned for the entire figure is 685.582764.

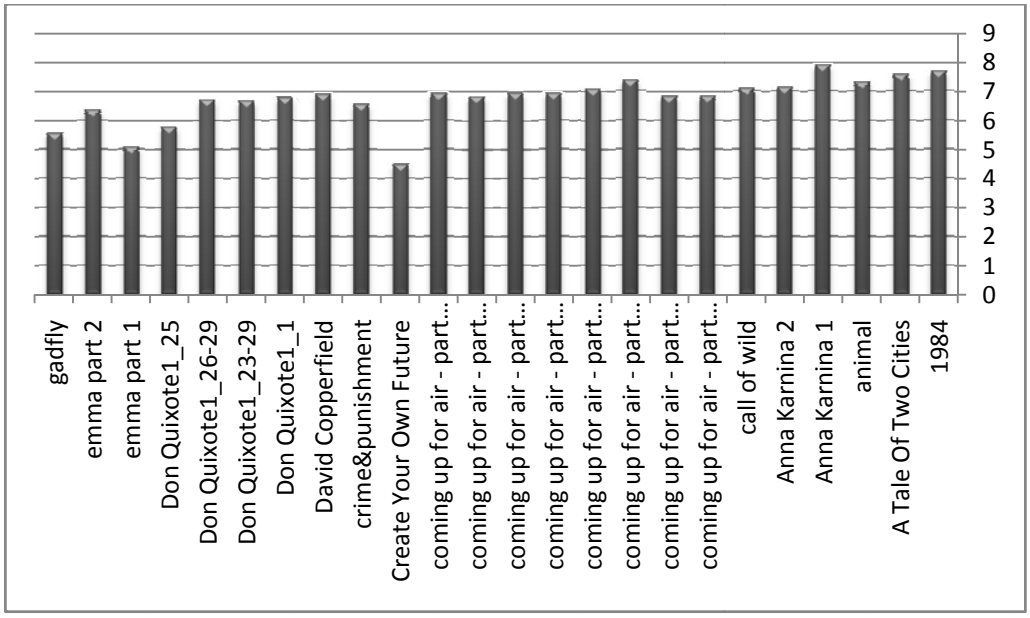


Figure 20: The percentage of words aligned for part of the figure

Other graphs are shown in Figures (21) and (22) for other features, the mean of the wrong characters and the marker hypothesis.

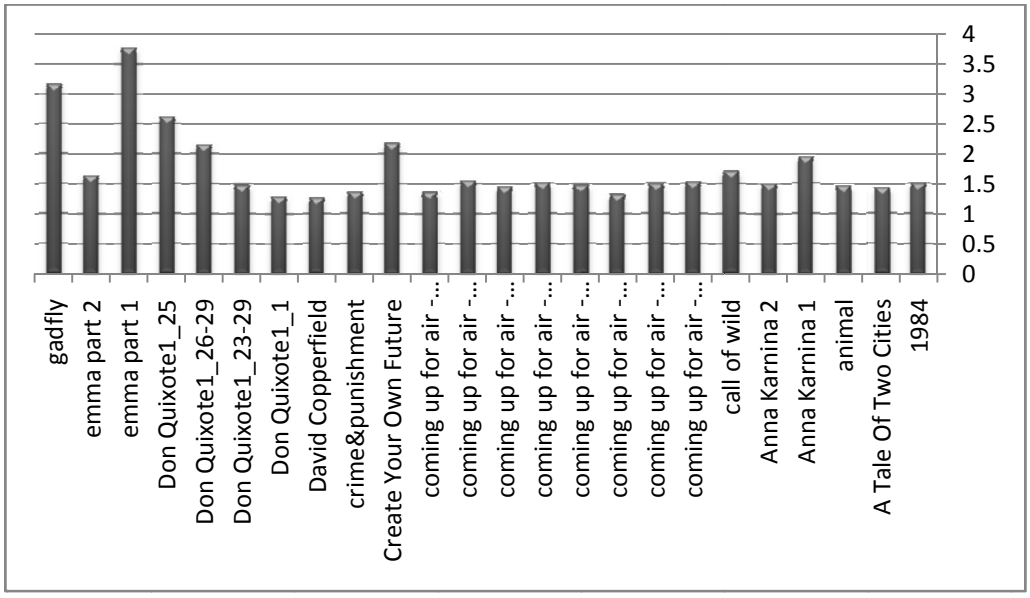


Figure 21: The sectional difference diagram in the marker hypothesis for multiple file formats

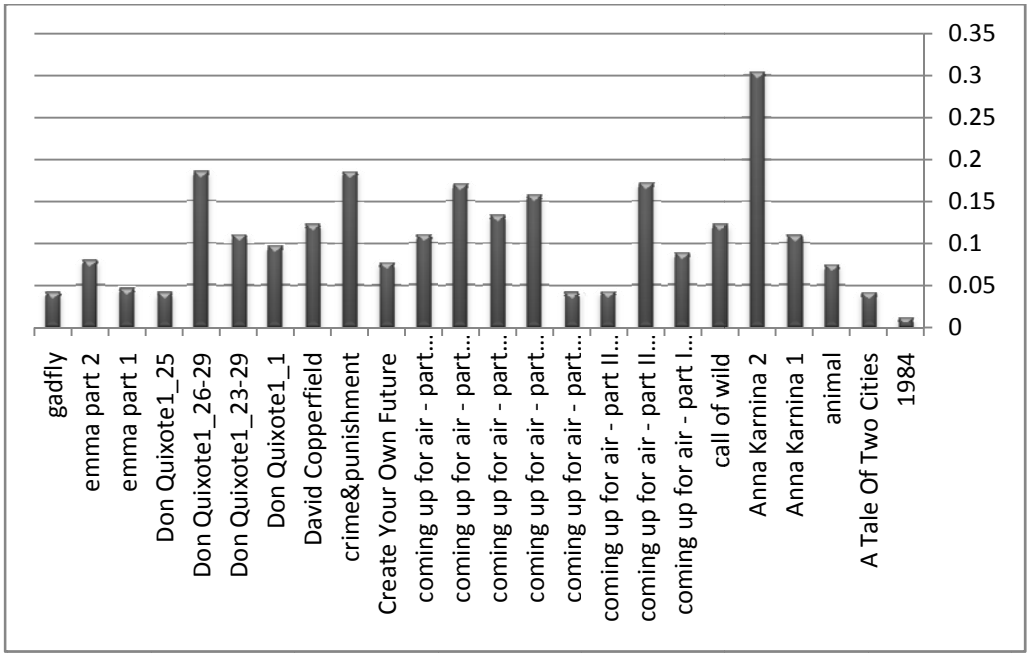


Figure 22: The average number of wrong characters in the files of the corpus

The average of incorrect characters in the whole body is equal to 119349808.0 and the mean difference between the sections obtained in the marker hypothesis is 703947559/1.

4.4. Final evaluation of the corpus

As can be seen in the figure, the final evaluation shows the figure of 6251408.72. This non-fuzzy number results from the rules of the fuzzy rule base based on the values of the six inputs generated. Finally, the membership grade is displayed for the output language terminology. As shown in figure (23), this figure has been rated good at 0.75 degrees and has been ranked as average.

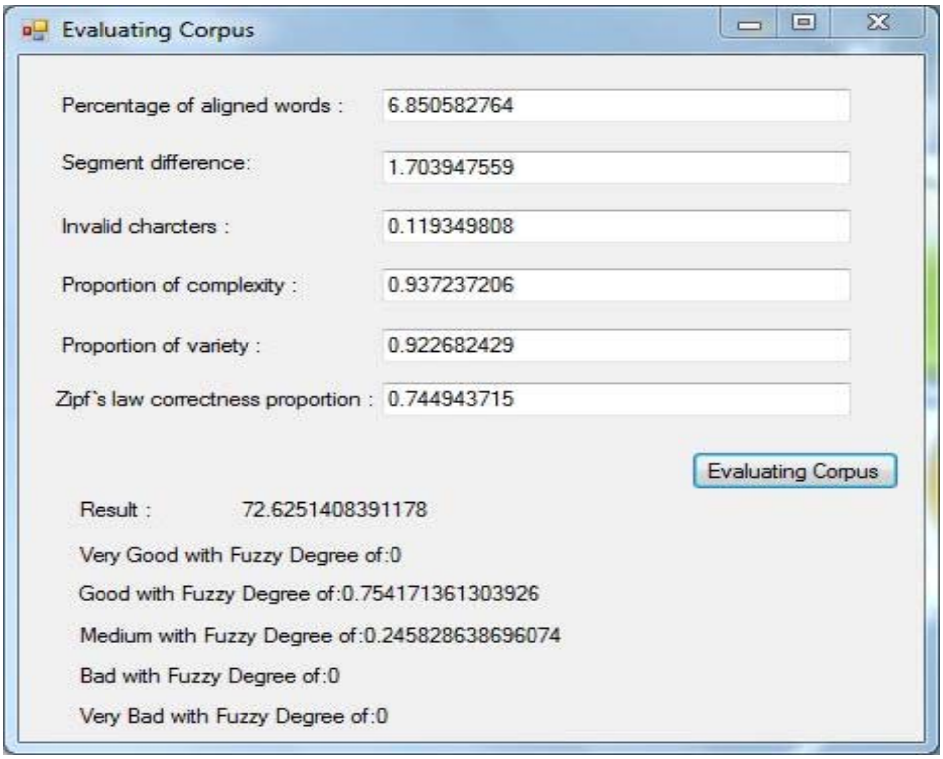


Figure 23: A sample of the final evaluation software

5. Conclusion

In this research, we try to evaluate the faces publicly, and finally, to examine the result of the method, we examine it on the figurative of one million English-Persian words. The final form used by the Persian-English corpus is at the sentence level of about one million words. This corpus is obtained by balancing the sentences of several books and translating them. The books that the statue is made of are for centuries, and include famous novels such as Oliver Twist, Harry Potter, Jane Eyre, Don Quixote, and more. This shape is aligned semi-automatically and is created in the form of several Xml files.

To evaluate the corpus, four main features of the figure, which are mentioned in references and books as representations of the corpus, are examined. These four characteristics are representativeness, reference standard, understandability for machine and limited size.

The last two qualities for bilinguals are definitely there. Therefore, we examined two features of standardization and representativeness of the corpus. For each of these properties, we selected and reviewed the features. For the standard of six linguistic features, good alignment, lack of wrong characters, proper separation, correct spelling, and correct punctuation were selected. For being representative, the complexity of the body, its diversity, and the correctness of the repetition of words in the figure were evaluated.

A way to quantify them was introduced or each of these features for each of these features,. A faithful translation into the text was calculated using the expectation of translating any word including English in Persian. For good alignment, the marker hypothesis was used. The absence of wrong characters was evaluated by examining the standard characters of each language in the sentences. Since spelling mistakes results in the lack of finding a translation for the word, we have included this feature in a faithful translation of the text. Due to the inappropriate separation of the marker hypothesis, this feature was also examined by the marker hypothesis. Finally, because texts were selected from edited books, the punctuation marks were not reviewed. We measured the complexity and diversity of the construct using the definitions available for these characteristics. For the correctness of the repetition of words, Ziff's law has been used.

After quantifying these six characteristics, we applied them to a part of the figure that had already been evaluated, with a parity of 6100 sentences. The results were used to create fuzzy language terms. Approximately 1,600 bases were created to evaluate the shape based on the language semantics of the six selected attributes. The rules were designed based on the average of input fuzzy semantics. Finally, software was developed for the final evaluation of the figure based on the six numbers, the fuzzy terms were defined and the created rules were designed. The final number obtained is a number between zero and one hundred.

The corpus used for evaluation found about 400,000 equally well-balanced sentences, resulting in a final evaluation of 72/62. Using the language terminology of the output, it can be said that the figure with a degree of membership of 0.75 is related to the good term language and with a degree of membership of 0.24 is related to the medium term language.

Of course, using this result, it cannot be said that the corpus is a good result in the machine translation. This result only indicates how well the entity is representative and standardized. In fact, this evaluation examines the quality of the entity, but does not examine the relationship of the entity with the translation machine. Meanwhile, it can be added that the Find-based translation software can be used as an aid to align sentences to build the corpus.

Other work that is proposed for future work is to compare these criteria with the output of the translation machine and to examine how much these qualities can improve the quality of the translation machine. To the extent that the author's knowledge helps, the criterion for finding the word translation (faithful translation) can be improved by the quality of the translation machine. As in other projects, this has been done to improve the translation machine. However, perhaps for other criteria this issue should be reviewed.

It is suggested that other criteria such as proper separation, correct spelling, and punctuation should be explored individually and complemented by the evaluation of the corpus.

References

[1] Tony McEnery, Richard Xiao, Yukio Tono, Corpus_based language studies: an advanced resource book. London & new york: Routledge Taylor & Francis Group, 2006.
[2] Tony McEnery, Andrew Wilson, Corpus linguistics.: Edinburgh University Press, 2001.
[3] Mostafa Assi, "From langue corpus to corpus Linguistics," Journal of Researchers (Journal of Social Sciences and Humanities Research) -Simple 8, 8, 14, 4, August and September, October and November 2006.
[4] Mostafa Assi, "A Plan for the Preparation of Computer-aided Specialized Cultures", in the Proceedings of the Second Conference on Theoretical and Applied Linguistics, Tehran, Allameh Tabatabaei University, 1994
[5] Graeme Kennedy, An Introduction to CorPus Linguistics. London: Longman, 1998.
[6] D. Santos , P. Rocha, "Evaluating CETEMPúblico, a free resource for Portuguese," in Proceedings of the 39th Annual Meeting on Association for Computational Linguistics,ACL '01, Morristown, NJ, USA, 2001, pp. 450–457.
[7] Bin LU, Tao JIANG,Kapo CHOW,Benjamin K. TSOU, "Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT," Workshop Programme, 2010.

[8] D Jurafsky, JH Martin, A Kehler, K Vander Linden, Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.: MIT Press, 2000.
[9] Langlais, P., Simard, M., Véronis, J., Armstrong, S.,Bonhomme, P., Debilil, F., Isabelle, P., Souissi, E.,Théron, P., "ARCADE: A cooperative research project on parallel text alignment evaluation," in Proceedings of the First International Conference on Language Resources and Evaluation, Grenada,Spain, 1998.
[10] L. Sun, S. Xue, W. Qu, X. Wang, and Y. Sun, "Constructing of a large-scale chinese-english parallel corpus," COLING '02: Proceedings of the 3rd workshop on Asian language resources and international standardization, pp. 1-8, 2002.
[11] S. F. Chen, "Aligning sentences in bilingual corpora using lexical information," Proceedings of the 31st annual meeting on Association for Computational Linguistics, pp. 1-8, 2002.
[12] K. T´oth, R. Farkas, and A. Kocsor, Sentence alignment of hungarian-english parallel corpora using a hybrid algorithm.: Acta Cybern, 2008, vol. 18.
[13] T. Talvensaari, J. Laurikkala, K. J¨arvelin, M. Juhola, and H. Keskustalo, "Creating and exploiting a comparable corpus in cross-language information," ACM Transactions on Information Systems (TOIS), February 2007.