DASTEX: a New Readability Formula based on Semantic Complexity

Mohammad Reza Besharati (Corresponding Author)

PhD Candidate, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, besharati@ce.sharif.edu

Mohammad Izadi, Associate Professor, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, izadi@sharif.edu

Abstract

Simple measures often couldn't count a deep complexity. In the case of semantic complexity, conventional readability formulas share a common style, a common sort of achievements and a common borders of limitation: These formulas lack a semantics-aware approach and as a result, a precise measurement of semantic complexity couldn't be done. In this paper, we introduce DASTEX, a novel semantics-aware complexity measure for semantic complexity of text. By DASTEX, a new layer of complexity analysis are opened for NLP, cognitive and computational tasks. This measure benefits from an intuitionistic underlying formal model which consider semantic as a lattice of intuitions. This yields to a well-defined definition for semantic of a text and its complexity. DASTEX is a practical analysis method upon this formal model. So a complete suite of idea, model and method are prepared to result in a simple but yet deep measure for semantic complexity of text. The evaluation of the proposed approach is done by 4 Experiments. The results show DASTEX is capable of measuring the semantic complexity of text in 6 application-tasks.

Keywords: Semantic Complexity, Semantics, Text Complexity, Readability Formulae

1- Introduction

Many readability formulas have been invented by researchers [1]. These formulas have a common aim: to evaluate the readability level or complexity score of a text. Such formulas numerically model the readability and calculate a rough value for it. Counting and enumeration of linguistic features are incorporated in them and as a result, these formulas usually follow a morphological computation of text elements. number of characters, Number of syllables, number of phrases, number of words, number of different part of speeches, number of sentences, number of ideas and concepts, number of named entities, number of relations and etc. are common micro-measures which together construct the macro-measures of conventional readability formulas (See figure 1 for a concise meta-model for conventional readability formulas).

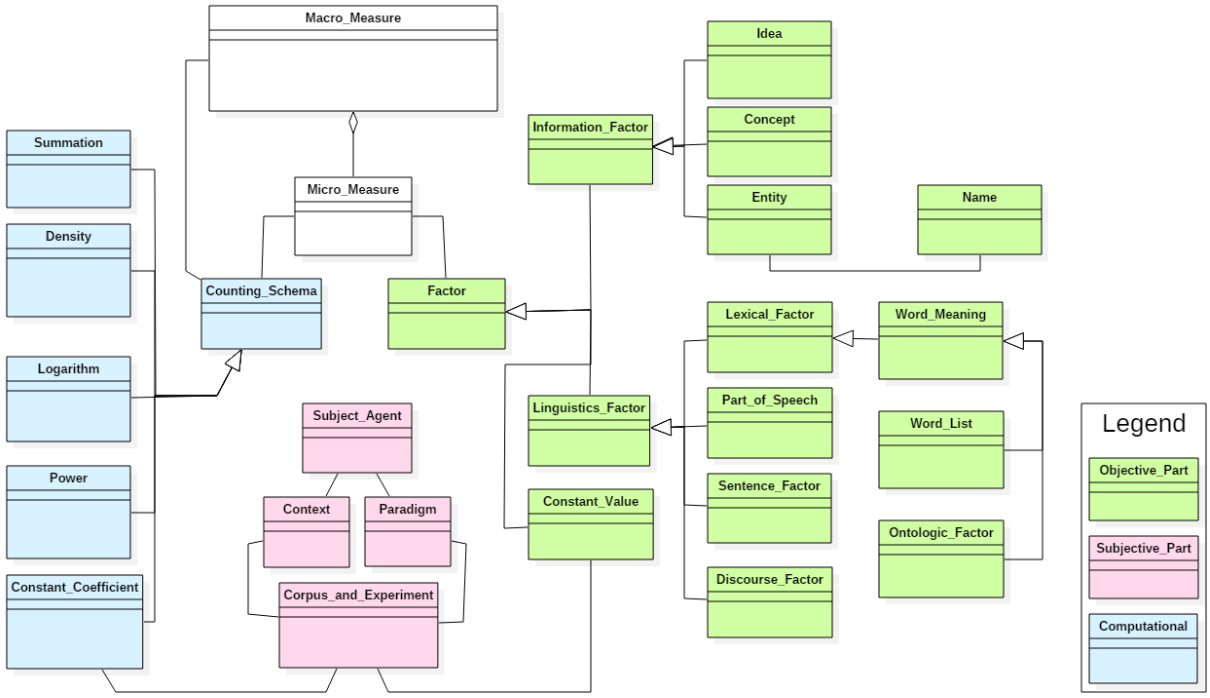


Figure 1- a concise meta-model for conventional readability formulas (The notation of UML Class Diagram is used for meta-modeling).

We could summarize the style of conventional readability formulas as an "element counting schema". This "sensing" oriented approach to complexity analysis yields to form-driven and text-style-biased measurements. Different texts from different authors with different contents which share a common text-style or rhetoric-form would result in similar readability scores. In spite of this sensing oriented approach, we could consider a new, alternative viewpoint: an intuitionistic approach to text complexity measurements.

Intuition, sensing, thinking and few others are different state of affairs or type of personality awareness in psychology [2]. What is the relation between psychology and

readability measurements? Each measurement is done under a psychological state of mental affairs (= cognitions). So different cognitive attitudes yield to different measurement paradigms. A sensing oriented approach counts the shallows, morphological elements. But an intuitionistic approach could consider more tacit, more diverse and deeper layers of a text. So for a better measurement of semantic complexity of text, we could prefer "intuition" over "sensing".

2- Related Works

In 19th century, there were worldwide independent efforts to match students with materials at appropriate levels of difficulty [11]. First modern readability formulas appeared in the 1920s (for a very early one, see [12]). By 1973, there were more than 200 various readability formulas [10].

Readability formulas usually rely on statistical processing and analytical results from a large collection of text documents [9], [13].

An essential underlying idea of almost all readability formulas is as this: obtaining an easy-to-compute proxy for semantic or syntactic readability of text [5]. This easy-oriented approach affects the attitude and scope of the invented formulas. Because of tacit and in-depth nature of semantics, easy-oriented approaches have has limited achievements in measuring semantic complexity of text. Even for measuring the overall reading difficulty of text passages, the conventional readability formulas are not good predictors [7], or at least they are inadequate predictors [8]. Measuring text complexity, without considering the text comprehension, results in not responsive estimates of text complexity [14].

The Applications could be listed as: Book leveling, suitability of reading materials for readers with different: ability levels [6], reading skills, ages, familiarity with background knowledge, language proficiency levels, Mental Mood, Psychologic state, Cognitive Health.

Formal semantics, such as operational semantics [20], is another sort of semantic modeling and a technique for semantic complexity could be based on it [21]. For example, coordination semantics could be served as a proxy for catching process meanings and complexities [22]. Simulation is another way for capturing operational process semantics [23]. Rule-based knowledge-aware quality definition approaches could also capture operational and non-operational semantics [24]. Based on the notion of Kolmogorov complexity [25], just after capturing the semantics, we could compute the semantic complexity by enumerating the involving building blocks.

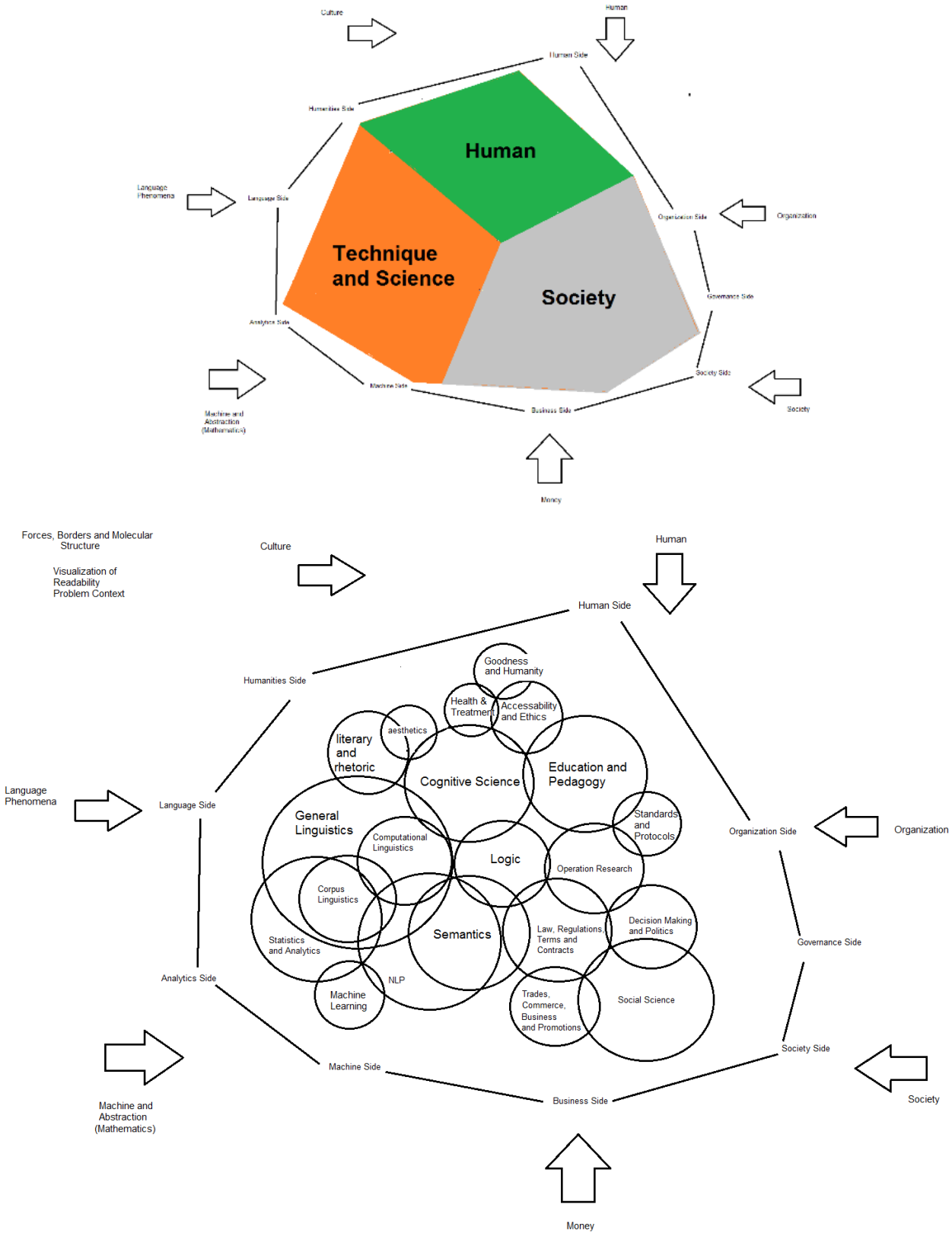


Figure 2- A Context Depiction for Text Readability and its Constituents Domains.

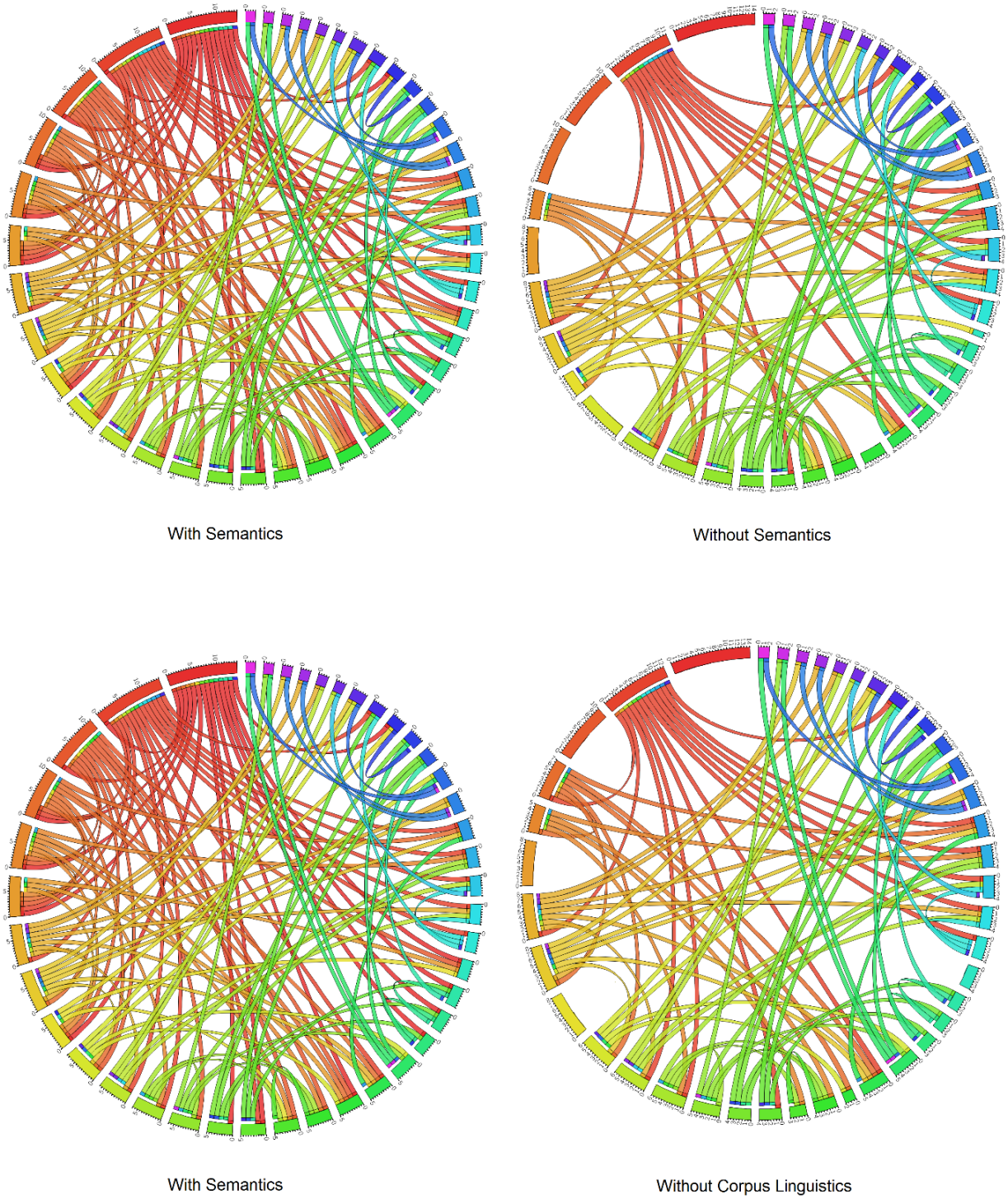


Figure 3- Intersections in the constituents domains of text readability context, in the Chord Diagramatic Layout [18].

3- The Proposed Model

In our computational approach to the Semantics, it is a process which is involving references (or mappings) from a content-system to entities, concepts, things, objects, states, realities, relations, sequences, scenarios and any other sort of "basic structural constituents". So a general model of "construction" can be used as a proper model for semantics and meanings. A hierarchy or a lattice could be a mathematical model for meanings. This model could be construct by using basic intuitions of subject understanding for each under semantics-study object (for a cognitive theoretical basis, see figure-4).

Two primary operation construct the meaning buildings: 1) putting the involving symbols (behalf of realities and entities) on the computation table, and 2) put them on each other to shape-up the overall hierarchical construction of the intended meaning. An example is depicted in the figure-5.

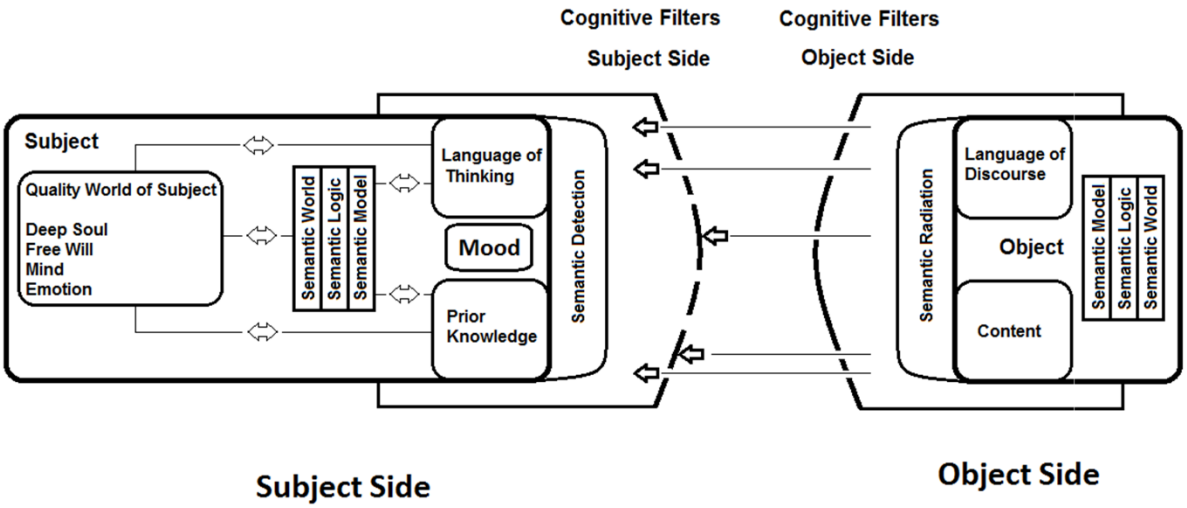


Figure 4- A Cognitive Theoretical Schema for Human Semantics Understanding

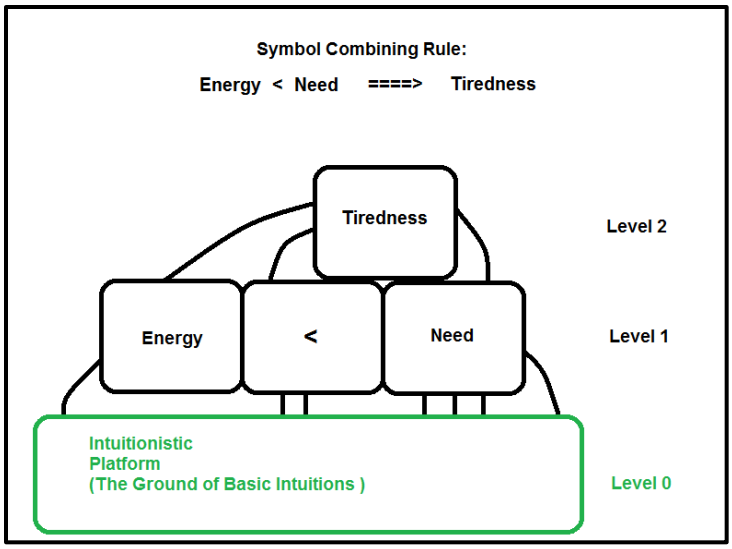


Figure5 - A symbolic Depiction of a Semantics Construction

Some meaning particles (for example some entities) are higher order and are made from a symbol construction process which take some lower order symbols as the inputs, then results some higher order symbols as the outputs.

So a formal model for semantics could be defined by these elements: 1) a set of primary symbols (or basic intuitions), and 2) a set of symbol combining rules. This is the essence of DAST model [15] for text semantics. DASTEX is based on DAST.

Each semantics theory should involves these statics and dynamics subsystems. So if we elaborate and detect these two primary parts, we could say "there is a semantics theory here". Each set of symbols (which are related to each other by some combining rules) could be considered as a Semantic Theory.

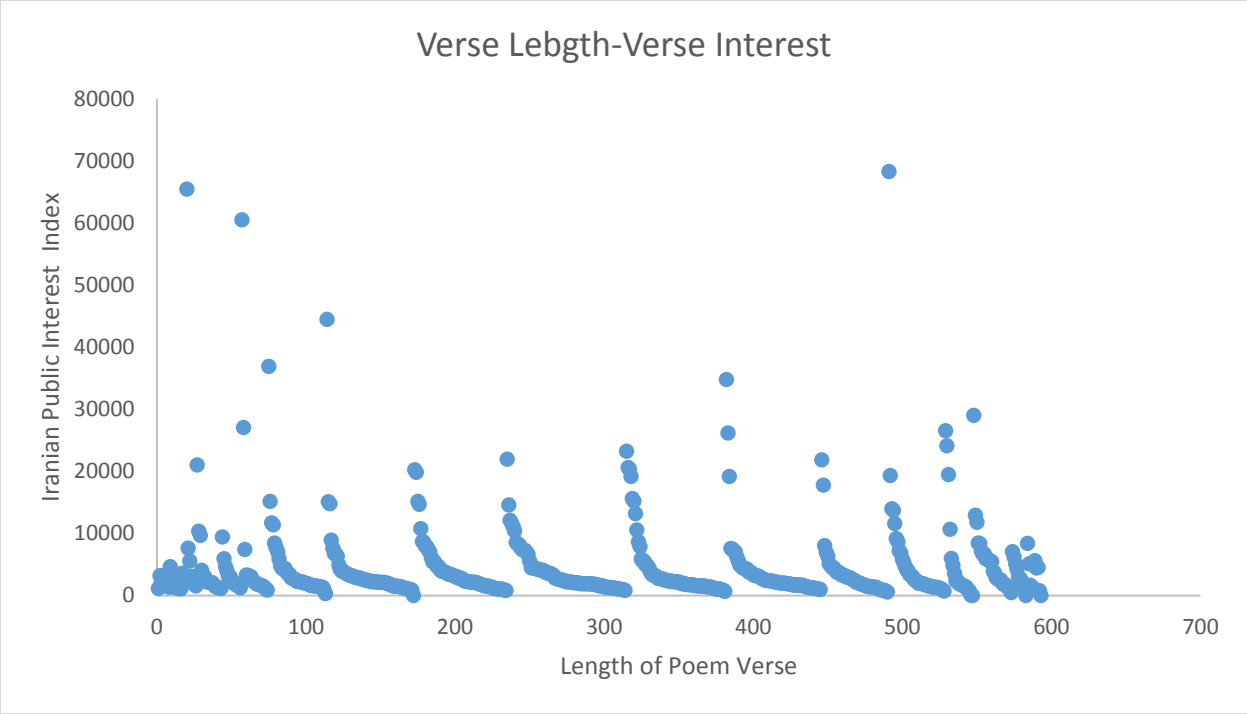
Definition 1. ∀T: text, DASTEX (T) = DAST Semantic Complexity Index for T

= Number of (Semantic_Theories (T))

4- Evaluation by Experiments

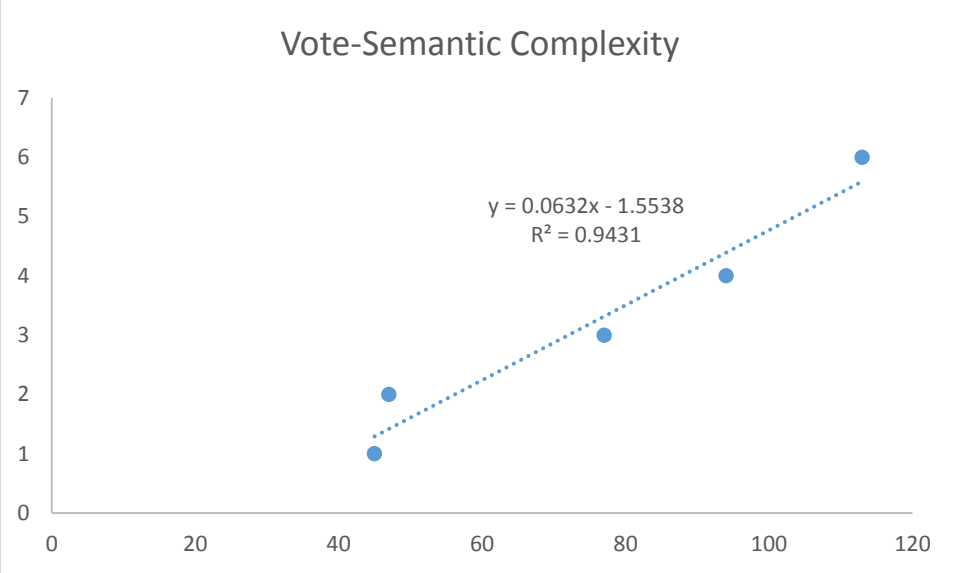
4-1- Experiment 1: Semantic Attraction Calculation

Hafez's poems are relatively popular in Iran. Based on the number of search results for each verse of Hafez poetry in a Persian web search engine, we tried to estimate the public interest index of each verse. The result is that there is a relationship between verse length (i.e. the semantic complexity capacity) and popularity.



4-2- Experiment 2: Semantic-Choice Calculation

We asked 376 individuals to choose a name (from a set of 5 name-choices) for a semantically complex book. There is a correlation between votes and semantic complexity of choices.



4-3- Experiment 3: Semantic Complexity Calculation

An important conclusion can be drawn from the DAST and DASTEX experiments, especially when the DAST results were matched between two Persian-speaking and Spanish-speaking statistical communities [19]: The semantic truth of meanings, like numbers, is a common human intuition. See figure 6.

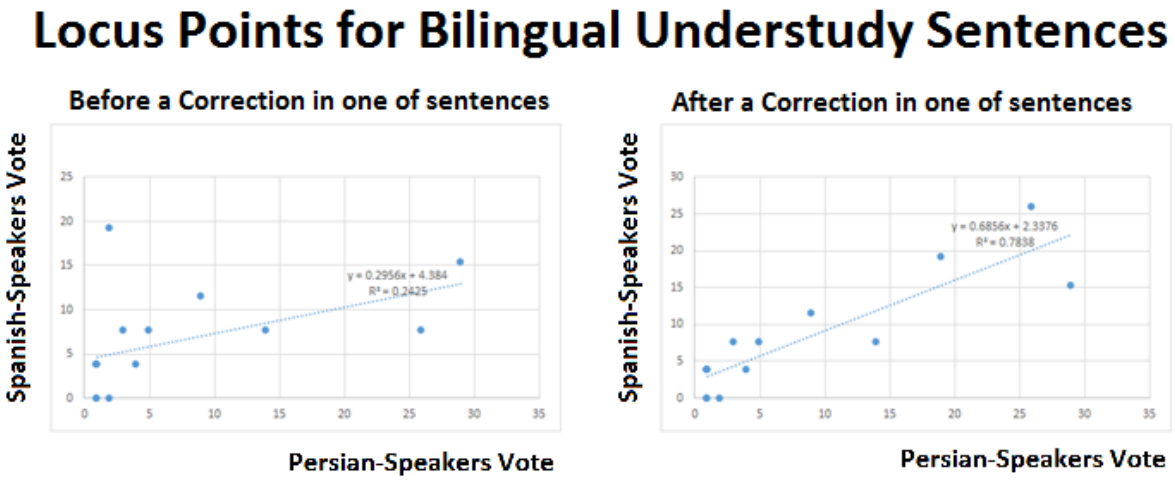


Figure 6- Semantic Complexity Votes for Bilingual Sentences.

In addition to objective logic and subjective logic, the "benchmark logic" must also be included in the semantic logic. With this, it will be possible to better judge about the semantic complexity. That is, for example, if a sentence was short and compared to a long sentence, then the effect of this length difference could be considered.

Some other results about DASTEX for a set of 80 understudy-sentences are provided in [15].

4-4- Experiment 4: Word Reputation Impact

The following data (figure-7) are related to the word choice questionnaire, which, as we see, has grown in three to four different phases from 0 to 2.5 (vote ratio between the first and second options). The difference is so great that it cannot be attributed solely to a change in the distribution of the input data stream.

Also, in almost three quarters of the process, the statistical community was members of a fixed telegram channel. It seems that "reputation" is one of the influential parameters in choosing a Persian word instead of its foreign equivalent. Because "reputation" is one of

the parameters of word selection, we see that the first option over time, with increasing its percentage, i.e. increasing its reputation, has had a steady growth in attracting the attention of the audience in semantic judgment.

A similar phenomenon is investigated in these papers: [16], [17].

Choices in Persian		Some of Data Value Points									
1	پرسش تکجو	5	9	12	17	19	20	21	22	24	25
2	پرسش از تک واقعیت	13	14	16	19	20	21	21	21	23	23
3	پرسش حقیقتی	1	1	1	1	1	1	1	1	1	1
4	پرسش واقعیتی	1	1	1	1	1	1	1	1	1	1
5	پرسش راستینه	2	2	2	3	3	3	3	3	3	3
Sum	مجموع	22	27	32	41	44	46	47	48	52	53

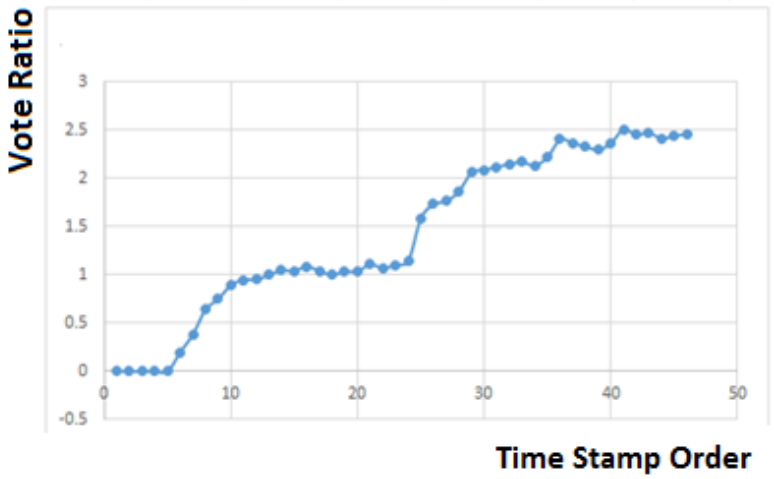


Figure 7- Experiment Results for Crowd-Based Persian-Word-Equivalent-Selection for "Factoid" word.

References

[1] Mühlenbock, Katarina Heimann. *I See what You Mean: Assessing Readability for Specific Target Groups*. University of Gothenburg, 2013.

[2] Making Sense of Software Development and Personality Types

[3] Mills, Chad. "Labeling and Automatically Identifying Basic-Level Categories." PhD diss., 2018.

[4] Islam, Zahurul, and Alexander Mehler. "Automatic readability classification of crowd-sourced data based on linguistic and information-theoretic features." *Computación y Sistemas* 17, no. 2 (2013): 113-123.

[5] Collins-Thompson, Kevyn. "Computational assessment of text readability: A survey of current and future research." *ITL-International Journal of Applied Linguistics* 165, no. 2 (2014): 97-135.

[6] Rush, R. Timothy. "Assessing Readability: Formulas and Alternatives." *The Reading Teacher* 39, no. 3 (1985): 274-283.

[7] Dalvean, Michael Coleman, and Galbadrakh Enkhbayar. "A New Text Readability Measure for Fiction Texts." *Available at SSRN 3097706* (2018).

[8] Ardoin, Scott P., Jessica C. Williams, Theodore J. Christ, Cynthia Klubnik, and Claire Wellborn. "Examining Readability Estimates' Predictions of Students' Oral Reading Rate: Spache, Lexile, and Forcast." *School Psychology Review* 39, no. 2 (2010).

[9] Collins-Thompson, Kevyn. "Computational assessment of text readability: A survey of current and future research." ITL-International Journal of Applied Linguistics 165, no. 2 (2014): 97-135.

[10] Fabian, Benjamin, Tatiana Ermakova, and Tino Lentz. "Large-scale readability analysis of privacy policies." In *Proceedings of the International Conference on Web Intelligence*, pp. 18-25. ACM, 2017.

[11] Humphreys, Alexandra H., and Jere T. Humphreys. "Reading difficulty levels of selected articles in the journal of research in music education and journal of historical research in music education." *Music Education Research International* 6 (2013): 15-25.

[12] Lively, Bertha A., and Sidney L. Pressey. "A method for measuring the vocabulary burden of textbooks." *Educational administration and supervision* 9, no. 389-398 (1923): 73.

[13] Anagnostou, Nikolaos K., and George RS Weir. "From corpus-based collocation frequencies to readability measure." In *ICT in the Analysis, Teaching and Learning of Languages, Preprints of the ICTATLL Workshop 2006*, pp. 33-46. 2006.

[14] Valencia, Sheila W., Karen K. Wixson, and P. David Pearson. "Putting text complexity in context: Refocusing on comprehension of complex text." *The Elementary School Journal* 115, no. 2 (2014): 270-289.

[15] Besharati, MohammadReza, and Mohammad Izadi. "DAST Model: Deciding About Semantic Complexity of a Text." *arXiv preprint arXiv:1908.09080* (2019).

[16] Sznajd-Weron, Katarzyna, and Jozef Sznajd. "Opinion evolution in closed community." *International Journal of Modern Physics C* 11, no. 06 (2000): 1157-1165.

[17] Hébert-Dufresne, Laurent, Samuel V. Scarpino, and Jean-Gabriel Young. "Macroscopic patterns of interacting contagions are indistinguishable from social reinforcement." *Nature Physics* 16, no. 4 (2020): 426-431.

[18] Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., … Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. Genome Research, 19(9), 1639–1645. https://doi.org/10.1101/gr.092759.109

[19] Besharati, M. R., & Izadi, M. (2019). DAST Dataset. https://doi.org/10.17632/2p7s6pb4vc

[20] Izadi, Mohammad, and Ali Movaghar Rahimabadi. "An equivalence based method for compositional verification of the linear temporal logic of constraint automata." *Electronic Notes in Theoretical Computer Science* 159 (2006): 171-186.

[21] Szymanik, J., & Thorne, C. (2017). Exploring the relation between semantic complexity and quantifier distribution in large corpora. *Language Sciences*, *60*(March), 80–93. https://doi.org/10.1016/j.langsci.2017.01.006

[22] Liaghat, Zainab, MohammadReza Besharati, Mohammad Izadi, and Ehsan Khamespanah. "Using Reo Formalism for Compliance Checking of Architecture Evolution with Evolutionary Rules." In *SoMeT*, pp. 725-738. 2019.

[23] Nowroozi, Alireza, Peyman Teymoori, Toktam Ramezanifarkhani, Mohammad Reza Besharati, and Mohammad Izadi. "A Crisis Situations Decision-Making Systems Software Development Process With Rescue Experiences." *IEEE Access* 8 (2020): 59599-59617.

[24] Besharati, Mohammad Reza, and Mohammad Izadi. "KARB Solution: Compliance to Quality by Rule Based Benchmarking." *arXiv preprint arXiv:2007.05874* (2020).

[25] Kolmogorov, Andrei N. "Three approaches to the quantitative definition of information'." *Problems of information transmission* 1, no. 1 (1965): 1-7.