Advances in Economics, Business and Management Research, volume 81
1st International Scientific Conference "Modern Management Trends and the Digital Economy: from Regional Development to Global Economic Growth" (MTDE 2019)

Linguistic corpuses as a digital means of comparative analyses of size adjectives (on the basis of English and Tatar)

Lutfullina G.F.
Kazan State Power Engineering University
Kazan, Russia
gflutfullina@mail.ru

Gilyazieva G.Z.
Kazan State Power Engineering University
Kazan, Russia
gilyazieva78@mail.ru

***Abstract* — The article is devoted to demonstrate how digital means are used in linguistic comparative analyses. The research is proved to be actual because size adjectives help to understand naïve language pictures of the world. The purpose of the article is to reveal differences in functioning size adjectives in English and Tatar. The methods of digital quantative analysis and a descriptive method were used. The sources for the collection and systematization of linguistic material were British National Corpus (https://www.english-corpora.org/bnc/) and Tatar National Corpus (http://web-corpora.net/TatarCorpus/search/). The research is carried out in line with those linguistic works where the idea of using statistic data is developed. The most frequent size adjective in Tatar is the adjective zur, the most frequent size adjective in English is the adjective little. The size adjectives thick / yuan have the same frequency in both languages. In Tatar the size adjectives ozyn and kechkene have the same frequency. In English size adjectives long-high-wide are distributed with different frequencies. When analyzing quantitative data using digital means of linguistic corpuses, it is necessary: first, to take into account the combining potential; secondly, to take into account synonyms.**

Keywords — size adjective, frequency, synonyms.

I. INTRODUCTION

Linguistic statistical indicators are provided by digital means. They help to determine word's frequency. This function allows to identify differences in synonyms' semantics, to establish contextual characteristics of synonyms and to distinguish between genre and stylistic features and meanings shades of lexical units. Digital means allow to select statistics of frequency analysis. Scientific hypothesis is precise and relies upon specific empirical data when we use such digital means as a linguistic corpus [1].

The corpus data serves as an experimental base to test hypotheses and to prove theories. Digital linguistic corpus allows to reveal statistical and linguistic patterns which are present in texts to create language models [2]. Size characteristics of objects are quantitative characteristics. The size variation of an object is expressed by a size adjective. In our opinion size characteristic is determined by the dominant parameter which dominates other parameters.

Size adjectives with indirect meaning dominate over size adjectives with direct meaning - *low cost, low income.*

One of the main tasks of our comparative analyses is to find differences in naive language pictures of the world where size adjectives are used to conceptualize real parameters.

According to Y. D. Apresyan linguistic tasks are to reveal the naive picture of the world from words' lexical meanings and to reflect the system interpretations. He considered Russian size adjectives *high, low.* The use of these adjectives is fully regulated by their following dictionary interpretations: *height = 'length of the object from the bottom to up', high = 'high in height', low = 'small in height'.* However naive geometry analysis shows that the language has a more complex system of rules to use size adjectives, to reflect the different features of meaning intuitively used in speech practice by native Russian speakers [6].

Object topological identification means to define object's specific spatial characteristics. There is a connection between the space and human three-dimension physiological and psychological type. Three-dimension space corresponds to three main coordinates of a human body: top-bottom, right side- left side, front and back.

I. Yu. Kuzina argues that vertical position is dominant in relation to horizontal, and frontal position is dominant in relation to the lateral. These relations are reflected in syntactic structures [7].

(1) *Then suddenly she saw him,* ***tall and thin and narrow*** *and neat* (KP5) [BNC] [8].

(2) *It was a* ***long narrow room*** *with large windows* (HMP) [BNC] [8].

According to I. Yu. Kuzina in order to characterize an object's size a person uses a coordinate system based on the measurement "top-bottom" (vertical position of the person), measurement "front-back" and "right-left" (asymmetry of the human body) [7].

II. METHODS AND MATERIALS

The aim of the article is to analyze size adjectives using digital methods of linguistic corpuses. During the research the methods of quantative analysis and a descriptive method were used. The source for the collection and systematization of linguistic material were the Britih National Corpus (https://corpus.byu.edu/bnc) (here and after referred as BNC) and Tatar National Corpus (http://web-corpora.net/TatarCorpus/search/) (here and after referred as

TNC). Total corpus of analyzed examples is six hundred examples.

III. DISCUSSION

Relative measurement naturally dominates in languages. Using size adjectives we qualify an object as having one or another indefinite size characteristic and we determine an object as *long / short, large / small*.

However there are noun groups of a special structure used to express the exact measurement.

According to Yu. D. Apresyan size parameter sometimes makes part of a word meaning:

1. Semantic component of a noun corresponds to semantic component of an adjective: *strong will = 'big + big ability... '=big ability…*'.
2. The conflict between a noun semantic component and an adjective semantic component may be neutralized *weak will = 'small + big ability...')*.
3. Nouns can express a neutral scale of a certain property [6].

Yu. D. Apresyan's concept corresponds to the modern theory of two scales: a scale of adjectives and a scale of nouns. Zhigue Xie examines functional applications of size adjectives *big / small* in English on the example of phrase *Big idiot*. The size adjective has an abstract meaning in this phrase while we analyze only direct meanings. However some interesting concepts concerning the adjectives are use in our research [5]. M. Morzycki develops the concept of" Bigness Generalization " [4]. He classifies size adjectives "*big", "enormous*" as positive parameter adjectives, while size adjectives *"small", "tiny"* are classified as negative parameter adjectives. He also develops the concept of a gradient noun which corresponds to the concept of a noun of variable size.

(3) *George is a(n) big/enormous/huge/colossal /mammoth/gargantuan idiot.*

(4) *George is a small / tiny / minuscule /Microsoft /diminutive/minute idiot.*

According to M. Morzycki only a combination *"big idiot*" definitely expresses a degree of idiocy that is contextually big. The formula states that the "*big idiot*" defines the variety of individual X whose magnitude of idiocy x corresponds to the contextual standard of "*being big*" and idiocy corresponds to the standard of "*being an idiot*". The opposite expression means that "*little idiot*" is an idiot whose idiocy meets the standard but whose "*little*" meets the standard of smallness. According to M. Morzycki [4] the scale of idiocy has a minimal element that is "*not idiotic at all."* It is necessary to take into account the opposite polarity between the scales of idiocy and the scale of smallness. According to M. Morzycki it is undesirable semantic emptiness responsible for the unpredictability of negative parameter adjectives characterizing the essential semantic component of "big". Moreover it is obvious that it does not depend on any contextual or pragmatic factors [4].

The article includes three parts with their own tasks:

1) To analyze frequencies indicators of size adjectives in both languages;
2) To investigate structures of noun groups including size adjectives and expressing accurate measurement;
3) To analyze words combinations expressing relative measurement.

IV. RESULTS

A. *Frequency indicators.*

Using digital means we carried out comparative analysis of size adjectives frequency in English and Tatar. Following data was revealed. The examples are taken from following linguistic corpuses (accessed 01.02.2019):

British National Corpus (96 263 399 tokens) [8];
Tatar National Corpus (26 000 000 tokens) [9];
Tatar Corpus "Tugan tel" (182 000 000 tokens) [10].

In order to normalize received statistical data the following formula is used. The frequency is calculated by the formula: frequency (words per million) = (number of words: number of words in the text) x 1000000

English word *"long"* – 55258 tokens.
Tatar word *"ozyn"* – 8 637 tokens.
Word "*long*" frequency = 55258: 96263 399*1000000= 574.
Word "*ozyn*" frequency = 8637: 26000000*1000000=332.

English word *"high"* – 37700 tokens.
Tatar word *"biek"* – 2 953 tokens.
Word frequency "*high"* = 37700: 96263 399 * 1000000= 391.
Word frequency "ozyn" =2 953: 26000000*1000000= 113.
English word *"wide"* – 11735 tokens
Tatar word *"kin"* – 8500 tokens.
Word "*wide"* frequency = 11735: 96263 399*1000000= 121.
Word "*kin"* frequency = 8500: 26000000*1000000=326.
English words *"thick*" – 4487 tokens; "*fat"* – 4381 tokens.
Tatar word *"zhuan"* – 30 tokens; "*yuan"* – 1246 tokens.
Word "*thick / fat*" frequency = 4487: 96263 399*1000000= 46
Word "*zhuan/yuan"* frequency = 1246: 26000000*1000000= 47
English words *"small"* – 42738 tokens; "*little"* – 61932 tokens.
Tatar word *"kechkene"* – 29068 tokens.
Word "*small*" frequency = 24382: 96263 399*1000000= 253.
Word "*little"* frequency = 61932: 96263 399*1000000= 643.
Word "*kechkene*" frequency = 8500: 26000000*1000000= 326.
English word *"big"* – 24382 tokens.
Tatar word *"zur"* – 212870 tokens.
Word "*big*" frequency = 24382: 96263 399 * 1000000= 253.
Word "*zur*" frequency = 40724: 26000000*1000000= 1566.

TABLE I. SIZE ADJECTIVES FREQUENCY

English		**Tatar**	
Long	574	Biek	332
High	391	Ozyn	113
Wide	121	Kin	326
Thick / Fat	46	Zhuan, Yuan	47
Little	643	Kechkene	326
Big	253	Zur	1566

Size adjectives frequency diagram.

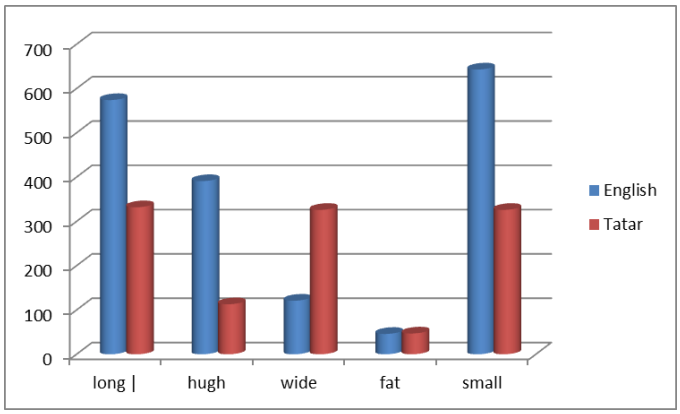

Fig. 1. Size adjectives frequency

It is concluded from the table and diagram data that in Tatar size adjective *zur* is the most frequent. In English frequency of the size adjective "*little*" slightly exceeds other size adjectives. We can state that size adjectives *thick / yuan* have the same frequency in both languages. In Tatar size adjectives *ozyn, kin, kechkene* have the same frequency. In English size adjectives don't have the same frequency indicators. Size adjectives *long – high – wide* are distributed in descending order with a significant interval. The obtained statistical data allows to predict the further research vector of compatibility potential.

(5) *…would take an entire evening to go out to Keighley. It's a bloody long way* (FM2) [BNC] [8].

(6) *Irteme-sonmy ul, ozyn yul uzyp, yuze omtylgan yarlarsyz, chiksez Dingezge kilep kushyla, yagni syyfat yuzgereshe ala (I. Valiullin. Mehebbet)* [TNC] [9].

(7) *…very similar building, a long sloping roof, not very high building. Isn't that, isn't that the second (unclear) isn't he, with* (KC4) [BNC] [8].

(8) *Tagyn shunysy bar, eger biekke, biek yort balkonyna menep zhirge karasam, minem sikeresem kile, nerseder tyelgysyz tarta — kosh syman kanat zheep, shul reveshle tyuben metelese kile. (Argy yar)* [TNC] [9].

(9) *The dark bulk of St Catherine's faced him on the other side of the wide road…* (B1X) [BNC] [8].

(10) *Kin yul sukmakka ejlende, sukmak tora-bara taraep yukka chykty. (R. Batulla. Karatay.)* [TNC] [9].

(11)*He was a thick man. Thick hair, thick eyebrows ,nose, lips, shoulders and tailoring…* (HA2) [BNC] [8].

(12) *Kinet tege shomly ishek achylyp, elege yuan keshe kassa aldyna bastyda, karchyga shikelle yutken kyuzleren egetler ostende yortep: — Kara, кara, galdish! (K. Tinchurin.)* [TNC] [9].

As we can conclude from given examples size adjectives compatibility in both languages is the same:

long way / ozyn yul [(5)], [(6)];

high building / bijek yort [(7)], [(8)];

wide road / kin yul [(9)], [(10)];

thick man / yuan keshe [(11)], [(12)].

It is concluded that the quantitative disproportion is due to different meanings. However we observe synonyms functioning – other adjectives assume size adjectives functions. Let's compare statistic data of the same words combinations in both languages:

English: *Little baby* – 71 tokens.

English: *Small baby* – 16 tokens.

Tatar: *Kechkene' bala* – 55 tokens.

(13) *Her old man was away fighting. They had a little baby, a girl. I was Fund of her, she reminded me of my.* (AE0) [BNC] [8].

(14) *Shul arany och kechkene bala belen Kavi duyrt konde Koch-hel belen uza aldy (Z. Zaynullin. Tatar ir-egetlere.)* [TNC] [9].

From given statistics you can state that in English one word combination dominates another one. The same is true for Tatar. Special attention is given to the synonym *belekej / little*. Taking into account the examples [(13)], [(14)] it is possible to say that the differences in functioning are insignificant and accepted as stable word combinations.

B. *Accurate measurement adjectives.*

Accurate measurement doesn't dominate in languages. Using numerals and size adjectives we qualify an object as having one or another definite size parameters.

Let us turn to the examples analysis. The aim was to make not only quantitative but also qualitative analysis of size adjectives structures and semantics. The quantitative representation is irrelevant.

Height

(15) *Racing down an active 2,380 ft (725,4 m) volcano at the second of 50 mph with only a board protection is considered by many thrill-seeking sports fanatics as the coolest sport* (KL6) [BNC] [8].

(16) *He used these sticks to build a 45-foot replica of Viking ship* (NF0) [BNC] [8].

In English a noun group can express exact measurement of the height parameter. This noun group usually includes an indefinite article and a noun [(15)], [(16)]. Between these obligatory components of a noun group there is a numeral and a measurement unit which define the noun *a 45-foot replica are placed* [(16)]. In example [(15)] an additional adjective is placed before the exact parameter *an active 2,380 ft (725,4 m) volcana.* The peculiarity of English is the use of national measurement units, for example *"feet"*, along with international measurement units. The rules of writing numbers require the use of a comma in thousand numbers. Size adjective "*high*" may be omitted. The height measurement is implied by noun's semantics. In Tatar similar noun groups are not found.

Length

(17) *At 6,500 feet (1981 meters) long, this zipline allows you to fly Superman at 93 miles per hour over gorgeous grasslands* (ACB*)* [BNC][8].

(18) *the Amazon rainforest gets its name from the 4,080-mile long Amazon river, the largest freshwater source on Earth* (CHK) [BNC] [8].

(19) *He bought a beautiful 40-foot yacht, with all the latest technical and safety gear, and had a pleasant voyage* (AD7) [BNC] [8].

(20) *Lesha 3 metrly ozyn baskych sojrep kilde. (Ya. Shafyjkov. Ber ochrashu – ber gomer)* [TNC] [9].

(21) *Mene chynlap ta 6 kvadrat metrly medpunkt bulyp chykty (L. Zahidullina. Boryndykta eshler hortime?)* [TNC] [9].

In order to express length English uses noun groups of the same structure:

Article / Demonstrative pronoun + Numeral + Measurement unit + Noun *meter* + Adjective *long* + Noun.

Examples: *this 6,500 feet (1981 meters) long zipline* [(17)]*; the 4,080-mile long Amazon river* [(18)].

In Tatar a noun group excludes an article. A measurement unit has a special affix -*ly*. A noun group is composed of:

Numeral + Measurement unit with the affix *ly* + Adjective *ozyn* + Noun

Examples: 3 *metrly ozyn baskych* [(20)].

In both languages size adjective *long / ozyn* may be omitted: *a beautiful 40-foot yacht* [(19)]; *6 kvadrat metrly medpunkt* [(21)].

Age

(22) *We wonder if the now 23-years-old girl who is studying cinema in Paris, will some day decide to follow her parents' example and take her kids back to experience, that she said* (GX9) [BNC] [8].

(23) *Luther Burbank, a famous American, developed more than 800 varieties of plants over his 55-year carrier* (AMW) [BNC] [8].

(24) *7 yellyk vakytta bary tik 20 men ekonomiya bire torgan 5 mashina gyna tormyshka kertelgen (A.Alish. Eserler)* [TNC] [9].

(25) *Utyz ellyk edebi stazhy bulgan, elle niche kitaby chykkan "Oly'" yazuchylar bar (G. Bashirov. 2012)* [TNC] [9].

(26) *Demobilizatciyalengen mayor Gavrilov hatyny, alty yashlek ugi kyzy belen Kazan kalasynnan utyz chakrymnar chamasynda ...kajtyp toshte (Z. Zajnullin. Tatar ir-egetlere)* [TNC] [9].

(27) *Heteren kalmasyn, egerme zhide yashlek kyzga – hette ul gyuzellernen gyuzele bulsa da – egerme ike yashlek eget ojlenmeyachek (F. Yarullin. Sajlanma ecerler)* [TNC] [9].

In English the dash is used to connect a numeral, a measurement unit and a size adjective *23-years-old girl* [(22)]. Size adjective is often omitted, for example *his 55-year carrier* [(23)]. Possessive pronoun determines the person. In Tatar a special affix -*lyk* is added to a temporal noun. The dash is not used *alty yashlek ugi kyzy* [(26)]. Size adjectives are sometimes omitted:

7 yellyk vakytta [(24)],

utyz ellyk edebi stazhy [(25)].

A possessive affix of a noun expresses the person. In example [(27)] a noun group has a special affix of case category *egerme zhide yashlek kyzga*.

Space.

(28) *Often called the lungs of the planet the 1.2 billion acre rainforest produces about 20 percent of the Earth oxygen* (CE9) [BNC] [8].

(29) *Ike katly selkenep tora torgan yort, ikenche katta chalysh byulmele sigez kvadrat metrly kvartir bulyp chykty ul (R. Batulla)* [TNC] [9].

(30) *Mene bu basu – totash dyurt yoz gektarly majdan (M. Mehdiev.)* [TNC] [9].

(31)*Yozerlegen gektarly basularny kim digende ike tapkyr boryngydan kalgan ysul belen, tepke bolgap eshkertep chygarga kirek (Kyzyl Gol: M. Malikova 2005)* [TNC] [9].

In both languages the above mentioned structures of noun group are used to express exact space parameter. The difference between languages is the use of different measurement units [(28)] - [(31)]. In Tatar the word *square* is also used *sigez kvadrat metrly kvadrat* [(29)].

C. *Relative measurement adjectives.*

In this article we consider English adjectives *big / small, little* and Tatar adjectives *zur / belekej, kechkene*. These adjectives usually characterize round objects. At the beginning of our article we stressed that the dominant parameter dominates another parameters. However the relative proportion of three parameters allows us to talk about the total volume or bigness. An object is characterized by its bigness when:

1) it differs by several parameters from normal size;

2) it has a vague shape and it is difficult to determine its visual space size;

1) it changes its volume but it conserves its proportion or shape, for example, round objects: *small / big watermelon*;

2) it has its shape but it is impossible to determine a dominant parameter, for example, *small / big cloud* [3].

Let's analyze and compare how these adjectives function in their direct meaning.

In British National Corpus

Adjective ‘*’big*’’ has 24382 tokens,
Words combination *big man* – 287 tokens [(32)].
Adjective ‘’*small’’* – 42738 tokens
Adjective ‘’*little’’* – 61932 tokens,
Words combination ‘’*little baby’’* [(33)] – 71 tokens,
Words combination *‘’small baby’’* [(34)] – 16 tokens.

(32)*... washing him? I don't go for that child, he's too big man* (KPE) [BNC] [8].

(33) *Her old man was away fighting. They had a little baby, girl. I was Fund of her, she reminded me of my* (AE0) [BNC] [8].

(34) *Martin had arrived at the gallery one morning, with his small baby in his arms, and the news that his wife had left him* (EFP) [BNC] [8].

In Tatar National Corpus

Adjective ‘’*zur’’* has 40724 tokens,
a words combination *zur keshe* – 249 tokens [(35)].
Adjective *kechkene* – 8500
Adjective *belekej* – 2462 tokens,
Words combination *kechkene bala* [(36)] – 37 tokens
Words combination *belekej bala* [(37)] – 15 tokens.

(35) *— Zur keshe disez inde alajsaa, - dide Golzhihan. Tokmachyn kise-kise (F. Yarullin. Sajlanma eserler)* [TNC] [9].

(36) *Shul arany och kechkene bala belen Kavi duyrt konde Koch-hel belen uza aldj (Z. Zaynullin. Tatar ir-egetlere.)* [TNC] [9].

(37) *Heterlisezder, belekej bala chagybyzda bezne "hedichettejnen ike bortege" dip yorteler ide (Eniki. Tash kalada)* [TNC] [9].

(38) *E etise nindi zur keshe (Z. Mahmudi.Serle kunak)* [TNC] [9].

V. Conclusion

A. *Frequency indicators.*

We can conclude that in Tatar size adjective *zur* is the most frequent adjective, in English size adjective *little* is the most frequent adjective. The size adjectives *thick / yuan* have the same frequency in both languages. In Tatar size adjectives

ozyn and *kechkene* have the same frequency. In English size adjectives *long-high-wide* are distributed with different frequencies. Analyzing quantitative data, it is necessary: first, to take into account the compatibility potential; second, to take into account synonymous phrases.

B. *Accurate measurement adjectives.*

On the basis of the analysis we can come to the following conclusions. In both languages there are similar structures (combinations of numerals, size adjectives and nouns) indicating the exact measurement of length, height, time. An accurate measurement involves a numeral and a measurement unit. Size adjective are represented in both languages, and noun semantics imply size parameter. A noun groups includes exact measurement unit. In English a noun group includes an article. In Tatar a noun or a measurement unit attaches special affixes. Languages use different measurement units. In Tatar Case category affects a whole noun group.

C. *Relative measurement adjectives*.

We can conclude that in Tatar size adjectives *of bigness* are less frequent than in English. The same is true for words combinations with similar meanings.

References

[1] P. Baker, C. Gabrielatos and T. McEnery, "Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word 'Muslim' in the British Press 1998–2009", Applied Linguistics, Volume 34, Issue 3, 1 July 2013, pp 255-278

[2] D.Biber, S.Conrad, Corpus Linguistics and Grammar Teaching, New York: Cambridge University Press, 1998.

[3] Ch. Kennedy, "Vagueness and grammar: The semantics of relative and absolute gradable adjectives", Linguistics and Philosophy, Volume 30, Issue 1, February 2007, pp 1–45.

[4] M. Morzycki, "Degree modification of gradable nouns: size adjectives and adnominal degree morphemes", Natural Language Semantics 17.2:175–203.

[5] Zhiguo Xie, "Where is the Standard? An Analysis of Size Adjectives as Degree Modifiers at the Semantic-Pragmatic Interface", Language and Linguistics, Vol. 15, Issue 4, 7 July, 2014.

[6] Yu.D. Apresyan, Issledovaniya po semantike i leksikografii/Research on semantics and lexicography/T. 1: Paradigmatika. M.: Yazy`ki slavyanskikh kul`tur, 2009, 568 s.

[7] [7] I.Yu. Kuzina, O parametrizaczii dejtsivtel`nosti yazy`kovy`mi sredstvami/Reality parameterization by language means//Vestnik Orlovskogo gosudarstvennogo universiteta, Seriya: Novy`e gumanitarny`e issledovaniya, 2011.# 1 (15). S. 161-165.

[8] [8] BNC – British National Corpus : https://www.english-corpora.org/bnc/

[9] [9] TNC- Tatar National Corpus: http:// http://web-corpora.net/TatarCorpus/search/

[10] [10]Tugan tel: http://tugantel.tatar/search.