Journal of Siberian Federal University. Humanities & Social Sciences
2020 13(3): 385–393

DOI: 10.17516/1997-1370-0561
УДК 811

Using Big Data Experiments in Cognitive and Linguo-Cultural Research in English and Russian

Olga A. Suleimanova and Inna M. Petrova
Moscow City University (MCU)
Moscow, Russian Federation

Received 05.06.2019, received in revised form 03.02.2020, accepted 10.03.2020

Abstract. Big data search instruments, Google and Yandex and others, offer new research opportunities, statistically very impressive. The questions arise here as to what extent we can trust the statistical data obtained, what kinds of hypotheses can be verified through Google and Yandex searching. The authors try to experimentally verify the hypothesis relating to the word order in the attributive group of the kind *healthy nourishing food vs nourishing healthy food* using Google and Yandex and, in this way, test the explanatory potential of the big data sources. The choice of word order is related to the theory of classes which explains cognitive mechanisms that govern the order of adjectives in a noun group. The empirical results obtained via Google and Yandex help to verify this hypothesis and identify cognitive grounds for choosing the order of attributes. The authors claim that data search engines, particularly, Russian Yandex and Google corpora, can help identify linguistically and culturally relevant concepts.

Keywords: Big data, cognitive principle, semantic experiment, theory of classes, word order of adjectives in the attributive group.

Research area: linguistics.

Citation: Suleimanova, O.A., Petrova, I.M. (2020). Using big data experiments in cognitive and linguo-cultural research in English and Russian. J. Sib. Fed. Univ. Humanit. Soc. Sci., 13(3), 385-393. DOI: 10.17516/1997-1370-0561.

* Corresponding author E-mail address: olgasoul@rambler.ru, miinna@yandex.ru
ORCID: 0000-0002-4972-0338 (Suleimanova); 0000-0003-3060-1387 (Petrova)

Introduction

The paper focuses on the issue of word order (WO) in the attributive group:

the authors will critically analyze current theories explaining WO (1);

put forward a hypothesis which accounts for WO (2);

support the theory with the experiment based on the big data (3);

analyze the potential and restrictions on the Google and Yandex experiments when exploring semantics (4).

1. Critical analysis of current WO theories

Linguists often claim that linguistics, as well as any other field of science, in its spiral progress keeps returning to "old" problems already seemingly extensively explored, but it approaches them from new, sometimes unexpected angles, discovering new features and trying new research methods and techniques. One of the problems which keeps linguistic minds excited is word order (WO) in attributive constructions, e.g. *healthy nourishing food* vs *nourishing healthy food*. This issue is also a challenge, teaching practice-wise.

Practitioners tend to rely either on their intuition and vast experience (if they are native-speakers of English, or any other language they are teaching), or on the self-explanatory potential of statistics. For instance, comprehensive research by S. Wulff (2003) reveals the so-called GSSSACPM rule (general-specific opinion-size-shape-age-color-provenance-material. However, the linguist claims that the rule is satisfactorily applicable to only 78% of the cases, while the remaining – statistically quite convincing, we have to admit – 22% are not covered by this rule. These 22% make it necessary to focus on the plausible explanation and to offer some convincing cognitive interpretation of the principles that govern the use of WO which the speaker opts for.

Other authors claim that deviations from the statistically preferred WO can be related to the individual choice, thus often leading to stylistically marked utterances, such as *great tawny-coloured intelligent eyes*. We believe that any stylistic idiosyncrasy, however, is based on some cognitive mechanism and should find a logical explanation which we shall offer below.

For instance, some authors suggest sequences: opinion – size – quality / character – age – shape – colour – participle – origin – material – type – purpose (Eastwood, 2002; Cohen & Lefebvre, 2005; Halliday & Matthiessen, 2014; Matthews, 2014). The sequence, unfortunately, can neither explain the equal acceptability of the word combinations *healthy nourishing food* and *nourishing healthy food*, nor the unacceptability of some of them.

Another approach is based on the statistically calculated principle: size – volume – softness – temperature – humidity – heaviness – form – age – colour (Ter-Minasova, 2008: 34) – *thick straight blond hair*. The adjectives denoting colour are located closest to the noun, while those denoting size are most distant: *little round green tables; big multicolored skirts; short thick blond hair; fat smooth round face.* Or in (Ter-Minasova, 2008: 229) the sequence is as follows: determiner – quality – size – length – form – colour – material – purpose: *a very valuable old gold watch; those smart brown snake skin shoes.*

The short list above demonstrates:

1) diversity of opinions related to the WO;

2) the terms used to define the semantic groups are not explicitly defined – it remains unclear what is meant by *quality*, or how *shape, form, length*. etc. are related;

3) variability of the type *a fat old lady* vs. *an old fat lady, or a wet dirty cloth* vs. *a dirty wet cloth* is not explained, though both of the options are acceptable.

It means that the above-mentioned theories lack sufficient explanatory potential as they are based on general and linguistic intuition rather than well-substantiated cognitive principle (Suleimanova & Petrova, 2018). We will try to combine the theoretical background with the research potential of a relative newcomer to the tool pool, investigation-wise, big data search systems.

2. Hypothesis accounting for WO

Further, we will define the key ideas, voiced by linguists, which will help us lay the

cognitive foundation of the theory we promote. We relate the approach to the theory of classes and their language categorization. Linguistic categorization has been in the focus of research attention, especially in the light of E. Rosch's works (1975; 1977, etc.) and in the papers to follow (Taylor, 2003; Cohen & Lefebvre, 2005; Tsohatzidis, 2014, etc.) where the authors claim that categorization as a cognitive idea is directly related to the physical survival of mankind in the hostile surroundings. To safely communicate with the unpredictable world, an individual has to be able to foresee and predict the properties of the objects s/he communicates with, s/he had to categorize the world, and each new object is immediately put into some category or class of objects, which makes communication successful, and leads to the individual feeling safe **(**Suleimanova & Petrova, **2018)**. So world categorization reflected in the language, as the projected world (Jackendoff, 1999), is of survival value.

What is relevant here is the degree of class definition, e.g., there are adjectives which define a stable (sub)class of objects that is easily empirically verified as the criteria defining the class are quite obvious and universally recognized. For example, in the word combination *a new beautiful nine-storeyed building* it is the attribute *nine-storeyed*. The word *new* also distinguishes a stable (sub) class, while the adjective *beautiful* introduces a vaguely defined class of objects as the criteria defining the boundaries of this class may essentially differ from person to person (Suleimanova, 1987). This difference leads to a hypothesis that it is along the boundaries of these subclasses the diversified variants can be expected – it can make the focus of special research in the future.

So cognitive categorization – with high probability – is reflected in the projected world, in particular, in WO of the attributive word combinations.

In line with the theory suggested above are similar views (though not enough explicit in some cases), such as that of M. Halliday, whose ideas rely on the theory of classes. He suggested four types of WO: Deictic, Numerative, Epithet and Classifier (Halliday & Matthiessen, 2014: 364). Although M. Halliday does not offer a general explanatory cognitive principle, not to mention the vagueness of the term "epithet"; he distinguishes classifiers, which can directly precede the noun. The "classifying" potential is shared by the names of the following categories: material, scale and scope, purpose and function, status and rank, origin, mode of operation as well as some other minor properties of the objects, forming their own categories (Halliday, Matthiessen, 2014: 377).

We share the view by N.A. Kobrina (2007) in that "the closest to the noun is a *classifying* adjective, as it expresses *constant, inherent properties and qualities* of the object. The next slot to the left is dedicated to explicative, a descriptive adjective..." (Kobrina, 2007: 127) (italics is ours – O.S, I.P.). As the linguist does not explain what is meant by the *classifying* feature (which, by the way, cannot be treated as self-explanatory), it is unclear which properties and under what conditions can be treated as *inherent* (see above).

Apparently many linguists tend to agree that the position closest to the noun belongs to words, denoting stable quality, more innate property, more intrinsic (Glucksberg & Danks, 1971), but it remains unspecified which qualities are stable, and which are not; what is more, why in some cases they are stable and, in a minute, they cease to be stable and are shifted from the stable position into a less stable one. It means that it is necessary to establish a clear link between the cognitive operations and their verbalization.

To the speaker it is relevant if there exists a class of objects he or she may refer to and how relevant it is for him or her to distinguish this particular class under the circumstances. The attribute which directly precedes the noun serves to denote such a class. For example, the Russian attributive group *узкая проселочная дорога* (narrow country road) implies that there exists a (sub)class of country roads, as opposed to highways; the property *narrow* is not treated as an absolute, universally recognized parameter and can be easily challenged: *It is not narrow – two cars can easily drive along*; *Какая же она узкая! Вполне можно разъехаться!*

In a neutral context *a fat old lady* is more probable as there exists a class of old ladies, besides the age is an extremely relevant characteristic. The word combination *an old fat lady* can occur in the contexts related to dieting, eating habits, obesity or fitness, where among the fat ladies the speaker singles out an old one. Similarly, choosing between word combinations of the type *white healthy teeth* vs *healthy white teeth* in the neutral context the speaker is likely to opt for *white healthy teeth,* as there exists an objectively identifiable class of people with healthy teeth (and this property is really important for people), and within this class the speaker distinguishes white-teethed people. On the contrary, the word combination *healthy white teeth* refers to white-teethed people (a less numerous class), and among them there are individuals with healthy or unhealthy teeth.

In other words, WO within the attributive group is defined by the relevance of some subclass for the speaker. The more typical the property is (cf. the statements in the theories analyzed above that the property is *stable, inherent, innate property*, actually defining the relevance), the closer it is to the noun; vice versa, the more subjective the assigned property, the farther it is from the noun. The speaker starts with the subjective feature to complete with the reference to the class: *a beautifully decorated tree-bedroom apartment*.

We have to say, though, that sometimes it is not very relevant for the speaker to prioritize between the properties, choosing one in favour of the other, as the properties may be equally relevant. For example, we can say in Russian *Я хочу найти себе красивого и умного мужа / умного и красивого мужа,* or in English *I want to marry some rich smart / smart rich guy,* where both features are equally important.

In other words, the principle which governs WO in the attributive group is related to the theory of classes and serves as the cognitive ground for the speakers' choice. The hypothesis based on the theory of classes was verified and the cognitive grounds for choosing WO were proved through the Google-based experiment (Suleimanova & Petrova, 2018).

3. Big data – based experiment

The suggested WO governing principle needed experimental verification. The original utterances were collected from the British National Corpus, the Russian examples were taken from the Russian National Corpus of the Russian Language (NKRJA); the WO in the attributive group was then changed and the resulting utterances were assessed as either acceptable or not. Earlier, linguists relied on the English native speakers who were asked to assess the utterances featuring such attributive constructions with regard to the acceptability of those utterances. Nowadays, linguists can rely on big data bases such as Google and Yandex and search for "hypothetical" word combinations (Suleimanova, 2019; Suleimanova & Demchenko, 2018; Petrova, 2018, 2019). If the search returns a compelling number of occurrences, it testifies to acceptability of the word combination in question. Such experimental procedure allows the authors to verify the hypothesis (see the detailed definition above), explaining the principles of WO, namely WO in the attributive word group is determined by the speakers' choice/focus on of the class of objects in the context.

We conducted both Google and Yandex-based search, which included a request for "competing" attributive phrases. There were 14 phrases: 7 Russian and 7 English with the following structure: attribute 1 + attribute 2 + attribute n + Noun. For example, in the attributive phrase *a beautifully decorated tree-bedroom apartment* the attribute *tree-bedroom* refers the noun *apartments* to a certain class of dwelling, while *beautifully decorated* attributes some subjectively identified feature to the apartment. When we change the WO in the phrase, we change the identification of the class by the speaker in terms of the features relevant for the speaker.

The results of the experiment can be seen in Table 1 which shows the results obtained via Google and Yandex for WO 1; Table 2 reflects the results for WO 2.

As our results indicate, WO 1 Phrases and WO 2 Phrases have different numbers of entries in Google and Yandex, but the relative ratio of values allows us to draw conclusions about the

Table 1. Occurrences of WO 1 in Google and Yandex (accessed 28/12/2018)

	WO1 Phrase	No. occurrences, Google Search	No. occurrences, Yandex Search
1	умная красивая девушка	662,000	80 000000
2	a smart beautiful girl	61,100,00	82 000000
3	большой красивый дом	2,280,000	121 000000
4	a big beautiful house	244,000,000	93 000000
5	умный сильный мужчина	258,000	50 000000
6	a smart strong man	148,000	80 000000
7	умный здоровый ребенок	539,000	77 000000
8	a smart healthy child	67,900,000	60 000000
9	вкусная здоровая пища	113,000	46 000000
10	delicious healthy food	75,900,000	43 000000
11	счастливая крепкая семья	238,000	42 000000
12	a happy strong family	189,000,000	53 000000
13	полная пожилая женщина	1,690,000	55 000000
14	a fat elderly woman	64,500,000	49 000000

Table 2. Occurrences of WO 2 in Google and Yandex (accessed 28/12/2018)

	WO 2 Phrase	No. occurrences, Google Search	No. occurrences, Yandex Search
1	красивая умная девушка	457,000	82 000000
2	a beautiful smart girl	54,800,000	79 000000
3	красивый большой дом	639,000	117 000000
4	a beautiful big house	182,000,000	96 000000
5	сильный умный мужчина	527,000	53 000000
6	a strong smart man	124,000,000	78 000000
7	здоровый умный ребенок	1,550,000	72 000000
8	a healthy smart child	24,100,000	57 000000
9	здоровая вкусная пища	189,000	46 000000
10	healthy delicious food	74,900,000	42 000000
11	крепкая счастливая семья	238,000	43 000000
12	a strong happy family	15,700,000	54 000000
13	пожилая полная женщина	1,670,000	0
14	an elderly fat woman	55,900,000	75 000000

preference of a particular phrase among the users. So, the experimental data contribute material for further analysis of WO of attributive phrases in Russian and in English and provide certain evidence of different tendencies in cultural perceptions and attitudes.

For example, the request for "competing" attributive phrases (no. 1-2) *умная красивая девушка* (a clever beautiful girl) and *красивая умная девушка* (a beautiful clever girl) reveals the following statistics – *умная красивая девушка* – 662.000 (Google)/80 000000 (Yan-

dex) entries, while *красивая умная девушка* – 457 000/82 000000 entries. Yandex does not provide a great difference in results, while Google is more specific. What follows is that the Russian cognitive practice admits existence of both clever, and beautiful girls, but higher frequency of the class of beautiful girls (if this figure is taken as 100%) against clever girls (70%) means that for the Russian linguistic picture of the world it is more natural to distinguish beautiful rather than clever girls. We can also conclude that this class is more culturally relevant for the Russian mentality.

The Russian word combinations *умный сильный мужчина and сильный умный мужчина* (a strong smart / smart strong guy) are related in a slightly different way: Russian mentality distinguishes a stable class of smart guys – 527 000 entries, and a less stable class of strong guys – only 258 000 entries, according to Google.

Comparing word combinations (we emphasize once again that different word orders are acceptable, and rigorously defined rules do not reflect the linguistic reality) can help define more "natural" classes, mirrored in the national cognitive practice. In this case, what is crucial is comparative data, rather than absolute figures, as taken separately, the figures would show only word combinations frequencies. As a research perspective we can suggest a comparative analysis of experimental data fished out of big data systems such as Yandex, Google, etc., which might provide unlimited material for linguistic and linguo-cultural research.

In the Russian cognitive experience, the child (no. 7-8) is seen in the same way as in the English language. A strong family (no. 11-12) in the UK (and the rest of the English-speaking world) is more important than a happy family (189 million vs 16 million entries), and in Russia the definitions are similar. The taste of food (no. 9-10) (189 thousand) in the Russian experience is more relevant than healthy food (113 thousand) – we have a more hedonistic vision than people in the English-speaking world, where these data are equivalent (75 and 76 million).

The search tools can also be quite instrumental in re-visiting the previous data. For example, in the doctoral theses by O.V. Afanasjeva (1994) devoted to the adjectival vocabulary, the author tackles the word order in the attributive word combinations and claims that in the combination, e.g. коричневый деревянный стол / *a brown wooden table* "out of many properties inherent to the object the speaker starts with choosing, as the most relevant, the indication to the material, thus relating the object some class <…> what concerns the word combination деревянный коричневый стол / *wooden brown table*, the analysis reveals that out of the class of brown tables the speaker singles out a sub-class of brown tables made of wood" (Afanasjeva, 1994: 322) (the translation is ours, O.S., I.P.). Though the author draws attention to variability in the WO, she explicitly claims that in English only one word order in the phrase *brown wooden table* is acceptable (Afanasjeva, 1994: 322). We tried searching through Yandex and found phrases *wooden brown table,* though in specific contexts – when pieces of furniture are discussed, that is why the colour is one of the key properties: *wooden brown table and blurred background, view from above, dishes on a wooden brown table.* In the last phrase, the crockery is referred to and how it is attenuated against the colour of the table. Then this phrase is widely used in the *antique furniture* slot.

When colour choice is in question, this WO is acceptable. We have a hunch that the cognitive principle we promote here may be treated as practically universal, or at least works for quite a few languages, and it might be quite a promising field of research.

4. Assessing the potential and restrictions on the Google and Yandex experiments when exploring semantics

The full potential of the search engines is not yet explored to full extent; we are facing only the beginning of such experimental procedures. Statistically, these tools offer compelling numbers as compared to the text corpora, e.g. COCA or BNC. The problem is that the search engines do not classify texts – the way texts are represented in the text corpora. Moreover, English texts in Google can be authored both by English native speakers and by those who speak English as their second/third lan-

guage, as English is spoken worldwide as the most popular international language. It means that English texts cannot be trusted linguistics-wise, and those systems are still to find ways to distinguish texts produced by different groups of people, for different purposes and in different styles. This state of affairs imposes strong restrictions on data in English.

As for Russian, either in Google or Yandex, the situation is different as Russian as a native tongue is spoken mainly in Russia and its former republics. What follows is that the Russian-based texts can be trusted, besides, the national mentality can be studied using both Yandex and Google Russian texts. The number of non-Russian native speakers contributing to these systems is negligible, statistics-wise.

Conclusion

The paper investigated the potential of **big data** experiments in cognitive and linguo-cultural research. The study reveals the potential of using search engines both as the source of reliable data and as the experimental tool that allows to verify the hypothesis explaining WO in the attributive phrase in cognitive terms. The hypothesis is based on the theory of classes, it admits WO variability in attributive groups, cf. *healthy nourishing food* vs *nourishing healthy food*. The choice of WO is related to the speaker's perception of the object and its relevance to him/ her. i.e. whether s/he distinguishes some (sub)class of objects s/he is further attributing some specific property to. The experiment allowed to distinguish more "natural" classes of objects projected onto the linguistic picture of the world. What remains unclear and calls for special investigation is that to define under what circumstances the speaker opts for this or that class is not possible through this experiment. Still, through analyzing attributive groups it is possible to draw conclusions about the mentality of the nation if we deal with the languages which are not spoken worldwide, such as Russian. At the moment the experimental potential of big data systems has not yet been extensively researched into and needs in-depth investigation.

References

Afanasjeva, O.V. (1994). *Ad''ektivnyj klass leksiki v sovremennom anglijskom yazyke i formy ego yazykovoj reprezentacii* [*Adjectival class of vocabulary in modern English and forms of its linguistic representation*]. Moscow, 395 p.

Cohen, H. & Lefebvre, C. (2005). *Handbook of Categorization in Cognitive Science*. Elsevier, 1136 p.

Eastwood, J. (2002). *Oxford guide to English grammar*. Oxford: Oxford University Press, 453 p.

Glucksberg, S. & Danks, J. (1971). Psychological scaling of adjective orders. In *Journal of Verbal Learning and Verbal Behavior*, 10 (1): 63-67.

Halliday, M.A.K., Matthiessen Christian, M.I.M. (2014). *Halliday's Introduction to Functional Grammar*. Fourth edition. Routledge, 805 p.

Jackendoff, R. (1999). *Semantics and Cognition*. The MIT Press, 8th. Ed., 283 p.

Kobrina, N.A., (2007). *Teoreticheskaya grammatika sovremennogo anglijskogo yazyka: Uchebnoe posobie* [*Theoretical grammar of modern English: textbook*]. Moscow: Higher school, 368 p.

Matthews, P.H. (2014). *The Positions of Adjectives in English*. Oxford University Press, 208 p.

Petrova, I.M. (2018). Kognitivnyj aspekt kombinatoriki sochinitel'nyh parnyh slovosochetanij v anglijskom i russkom yazykah [Cognitive aspect of combinatorics of compositional paired phrases in English and Russian]. In *Modern science: actual problems of theory and practice*, 9, 151-156.

Petrova, I.M. (2019). Kombinatorika binominal'nyh konstrukcij kak otrazhenie processa ikonicheskoj reprezentacii ob''ektov dejstvitel'nosti [Combinatorics of binomial phrases as a reflection of the process of iconic representation of objects of reality]. In *Cognitive studies of language*, 37, 621-625.

Rosch, E. (1975). The nature of mental codes for color categories. In *Journal of Experimental Psychology. Human Perception and Performance*, 1, 303-322.

Rosch, E. (1975). Universals and cultural specifics inhuman categorization. In *R. Brislin, S. Bochner, & W. Lonner (Eds.), Cross-cultural perspectives on learning*. New York: Halsted Press, 305-320.

Rosch, E. (1977). Human categorization. In N. Warren (Ed.). In *Advances in cross-cultural psychology*, Academic Press 1, 1-72.

Suleimanova, O.A. (1987). *Nekotorye semanticheskie tipy substantivov i ih aktualizatory ves'/celyj i all/whole* [*Some semantic types of substantives and their actualizers* весь/целый *and all/whole*]. Candidate dissertation, Moscow, 189 p.

Suleimanova, O.A. & Petrova, I.M. (2018). Eksplanatornyj potencial teorii klassov dlya lingvisticheskogo issledovaniya: poryadok sledovaniya opredelenij [Explanatory potential of the theory of classes for linguistic research: the order of elements in an attributive phrase]. In *Philology: Research*, 3, 52-64.

Sulejmanova, O.A., Demchenko, V.V. (2018) Ispol'zovanie BIGDATA v eksperimental'nyh lingvokognitivnyh issledovaniyah: analiz semanticheskoj struktury glagola shudder [The use of *BIGDATA* in experimental cognitive studies: analysis of the semantic structure of the verb shudder]. In *Cognitive studies of language*, 33, 466-472.

Suleimanova, O.A. (2019). Cemanticheskij eksperiment: novye vozmozhnosti v sisteme koordinat big data [Semantic experiment via big data system]. In *Cognitive studies of language,* 36, 427-432.

Taylor, J.R. (2003). *Linguistic Categorization*. *Prototypes in Linguistic Theory.* 3rd ed., Oxford: Oxford University Press, 312 p.

Ter-Minasova, S.G. (2008). *Yazyk i mezhkul'turnaya kommunikaciya: Uchebnoe posobie dlya studentov, aspirantov i soiskatelej po special'nosti «Lingvistika i mezhkul'turnaya kommunikaciya»* [*Language and intercultural communication: a textbook for students, postgraduates and applicants majoring in Linguistics and intercultural communication*]. Moscow: Word, 261p.

Tsohatzidis, S.L., ed. (2014). *Meanings and Prototypes: Studies in Linguistic Categorization*. Routledge, 584 p.

Wulff, S. (2003). A multifactorial corpus analysis of adjective order in English. In *International Journal of Corpus Linguistics*, 8(2), 245-282.

Использование больших данных в когнитивных и лингвокультурологических исследованиях английского и русского языков

О.А. Сулейманова, И.М. Петрова
Московский городской педагогический университет (МГПУ)
Российская Федерация, Москва

Аннотация. Инструменты получения больших данных, Google, Яндекс и другие, предоставляют статистику, которая открывает новые возможности для исследователей. Это, в свою очередь, порождает вопросы о надежности таких данных, а также о том, какие гипотезы могут быть проверены с помощью поиска в Google и Яндексе. Авторы экспериментально проверяют гипотезу, связанную с порядком слов в атрибутивной группе типа *healthy nourishing food vs nourishing healthy food*, используя Google и Яндекс, и, таким образом, подтверждают объяснительный потенциал данных инструментариев. Выбор порядка следования атрибутов в группе связан с теорией классов, которая объясняет когнитивные механизмы, управляющие очередностью расположения прилагательных. Эмпирические результаты, полученные с помощью Google и Яндекса, позволяют подтвердить эту гипотезу и выявить когнитивные основания для выбора порядка следования атрибутов. Авторы утверждают, что поисковые системы, в частности русскоязычный Яндекс и англоязычный Google, могут быть полезны при исследовании языковых и культурно значимых концептов.

Ключевые слова: большие данные, когнитивный принцип, семантический эксперимент, теория классов, порядок слов прилагательных в атрибутивной группе.

Научная специальность: 10.02.00 – языкознание.