ЛИНГВИСТИЧЕСКИЕ ПАРАМЕТРЫ КОНЦЕПТОВ

LIST И CATALOG: ВARIANT OБРАБОТКИ ЯЗЫКА
ДЛЯ КОМПЬЮТЕРНЫХ СИСТЕМ

Андрей Валерьевич Носов
старший преподаватель кафедры иностранных языков и связей с общественностью
Пермский национальный исследовательский политехнический университет

В статье рассматриваются два близких по значению понятия, отражающие мировоззрение человека. Целью работы является определение наиболее подходящего концепта для его дальнейшего применения в качестве упорядочивающей или перечислительной единицы в семантических исследованиях. Эти единицы существуют независимо от наблюдателя и могут проявляться в системе различий и противопоставлений. С учетом нашего предположения, выделяем LIST и CATALOG как близкие по значению слова с целью выявления факторов, определяющих выбор того или иного концепта, исходя из объективных семантических критериев. С этой целью проводится предварительная разработка схемы логистического регрессионного анализа. В центре внимания оказывается выявление содержания и динамики понятий LIST и CATALOG с позиций (1) четырех стилевых регистров языка (академического, разговорного, литературного, газетного), (2) семантического признака притяжательного значения, (3) семантического признака единственного/множественного числа, (4) постмодификаций, (5) премодификаций и (6) конструкций родительного падежа.

Ключевые слова: компьютерная лингвистика; логистическая регрессия; сравнительный анализ; семантика; близкие по значению понятия.
1. Collocation analysis;
2. Definition of context register (style);
3. Grammatical pattern differentiation;
4. Dialect usage;
5. Referential analysis (connotations).

From this point of view a look at English dictionary definitions detects rather a few similarities in meaning between LIST, CATALOG, REGISTER, SCHEDULE, NOMENCLATURE, ROLL and INVENTORY. In fact, in certain cases the definitions of these lexical items seem to be circular. Considering these notions within a computational approach reduces the collection to two samples. Hence the most accurately allocated opposition is LIST and CATALOG. If we consider the definitions of LIST taken from several dictionaries (New Shorter Oxford English Dictionary (1993), Collins English Dictionary (1994), Collins Thesaurus of the English Language (2002), Longman Dictionary of Contemporary English (2009), Random House Webster’s College Dictionary (1996), Ozhegov S. I. and Shvedova N. Yu Defining Dictionary of Russian Language (1992) and others) it becomes evident that REGISTER, SCHEDULE, NOMENCLATURE, INVENTORY, ROLL almost infallibly occur as one of the variations of LIST; at the same time CATALOG also has a strong resemblance to LIST, that U. Eco in his work “Infinity of Lists” [Eco 2009] often substitutes one for another. Thus I put a task to find out the main lines of these two concepts.

Firstly I give their definitions below:

LIST
1. A series of names or other items written or printed together in a meaningful grouping or sequence so as to constitute a record;
2. Computing a linearly ordered data structure;
3. A database containing an ordered array of items (names or topics);
4. Item, point;
5. An alphabetical index of names and topics along with page numbers where they are discussed.

CATALOG
1. A list or itemized display, as of titles, course offerings, or articles for exhibition or sale, usually including descriptive information or illustrations, a publication, such as a book or pamphlet, containing such a list or display;
2. An enumeration;
3. A card of the contents of a library or a group of libraries, arranged according to any of various systems;
4. Any record;
5. A written work or composition that has been published (printed on pages bound together);
6. A complete list of things; usually arranged systematically;
7. A series, as of names or words, printed or written down.

By comparison of these dictionary definitions:
1. We note that the words LIST and CATALOG in modern English are poly-semantic (in the analysis I take into consideration only nucleus dictionary definitions which allow me to define their values in modern English; I don’t consider obsolete definitions or others used for special purposes).
2. We appreciate the componential structure containing rather similar integrated and differential components. The Presence of integrated components in the structure of concepts under consideration is reflected by five groups of semantic elements:
   1. Series, set, sequence, grouping, array;
   2. Things, names, numbers, items, members, words, topics, files, point;
   3. Data structure, database, record;
   4. Written, printed, imagined, ordered, contributed and stored;
   5. Computing, containing, listing.

At first glimpse the differential components in the structure of definitions reflected by verbal or non-verbal symbolic representations are the same. Eventually, it isn't difficult to find out some utterances where these concepts can replace each other:

   (1) Rather than lump genomics and genetically modified organisms (GMO) into the upcoming chapter where I elaborated the list of technologies that will transform agriculture, I chose to address them here because a topic so controversial and misunderstood warrants extra time to sort through.

   (2) I have read every historical work that I have been able to lay my hands on, from a catalog of dry facts and dryer dates to Green’s impartial, picturesque “History of the English People”; from Freeman’s “History of Europe” to Emerton’s “Middle Age”.

Elaborating the similarities in the usage of these words and in their context dependencies we seek to show a number of distinctions at the semantic level. At this stage we put forward the first hypothesis. In spite of the fact that CATALOG has more dictionary definitions (i.e. more concrete meaning) its scope is much narrower than the LIST one.

(3) «... un caso de enumeración caótica, donde aparentemente no hay ningún nexo entre los varios elementos de la lista. No obstante, merece la pena desembozar una distinción entre enumeración conjuntiva y enumeración disyuntiva. Una enumeración conjuntiva reúne también cosas distintas que aportan al conjunto una coherencia, ya que es el sujeto mismo el que las ve o están consideradas en un idéntico contexto; por el contrario, la enumeración disyuntiva expresa una fragmentación, una especie de esquizofrenia del sujeto que concibe una secuencia de impresiones.
disparatadas sin conseguir atribuirles utilidad ninguna» [Eco 2009: 2]. This judgment underlines free and open character of the category LIST.

In the sentence below the semantics of LIST includes the meaning of CATALOG and transforms it into one of its features. Hence the sum of tokens LIST should be greater than the CATALOG one in the same corpus.

(4) El catálogo de un museo representa un ejemplo de lista práctica, que se refiere a objetos existentes en un lugar determinado, y como tal está necesariamente acabada (Translated by the author from Spanish: “The catalog of a museum represents an example of practical list, which refers to existing objects in certain place, and as such it is closed”) [Eco 2009: 20].

The second hypothesis. We reveal CATALOG in special contexts of fiction, news and spoken register too, but it has respectively low frequency. Hence I suppose the application of CATALOG is appropriate in more academic context whereas LIST is introduced in more universal way and is equally represented in all four registers. I will try to extract it from the data and theory focusing on (a) the distribution of LIST and CATALOG across four language registers mentioned above.

The third hypothesis considers the academic register as equally appropriate for both LIST and CATALOG.

(5) For a list of Fourier’s publications see the Catalog of Scientific Papers of the Royal Society of London.

One more criterion we use of in order to distinguish LIST and CATALOG are grammatical patterns. Fairly often near-synonymous words don’t have the similar grammatical patterns. Probably it is the most frequently corpus-linguistically studied area. In my research I come across four semantic features (possessive meaning, pre-modification, post-modification and genitive meaning).

The semantic feature of (b) possessive meaning that I analyze arises in this case from narrowing semantics and uniform effects of the notions LIST and CATALOG. R. Barker says that quantification property, termed possessive existential import, is intimately related to the notion of narrowing [Barker 1995], i.e. narrowing has implications for compositional analysis of any meaning. I apply these semantics to the issue of definiteness extent of LIST and CATALOG. Here I suppose LIST tends more than CATALOG to expanding semantics and not to narrowing (the fourth hypothesis).

One linguistic tool that reflects the semantic development level of the words is the analysis of the structural devices used to modify noun phrases. In English, modifiers stand before the head noun – (c) pre-modifiers – or after the head noun – post-modifiers. Most pre-modifiers in English are phrasal. We allocate three main structural variants of pre-modifiers: nouns (donor list), participial adjectives (detecting list) and attributive adjectives (short list). The Difference in the level of interaction between pre-modifiers and head nouns LIST and CATALOG suggests their order of semantic covering and introspective degree. We assume LIST as more pre-modified than CATALOG (the fifth hypothesis).

In contrast, (d) post-modifiers indicate concept functionality degree i.e. show which concept has a larger functional area. We define two groups of post-modifiers:

1. clausal
   – finite relative clauses (catalog indexes that represented this topic)
   – non-finite participial clauses (a. list tracking the polls; b. a new catalog held in position)
   – to-clauses (list to modify)
2. phrasal
   – prepositional phrases (‘Fashion’, new catalog)
   – appositive noun phrases (list of requirements)

Pre-modifiers and post-modifiers are more common for informational written registers (fiction, newspaper or academic) than for others [Varantola 1984; Halliday 1987]. Generally, they are about equally common [Biber 2002: 578]. In this work, I admit that LIST being more general concept than CATALOG needs to be specified by operations, caused by it. I allow the hypothesis about LIST as more post-modified than the CATALOG in all the registers to extract from the data and theory (the sixth hypothesis).

In the present study we also introduce (e) the genitive meaning predictor (of-genitives and ‘s-genitives). I introduce some restrictions at this stage. Firstly, of-genitives and ‘s-genitives are considered as two roughly equivalent ways of saying the same thing [Labov 1972]. Secondly, “have sentences cannot account for the wide range of semantic readings available in many’s genitive constructions”. [Shuemaker 1975: 71] Hence we exclude this point from the analysis. Thirdly, we restrict our attention to genitive occurrences with non-pronominal possessor/possessums.

From this point of view the genitive meaning complements tenors of concepts in the ambiguous realizations, as well as their measuring and explanatory features based on the principle of similarity.

Our aim is to test two assumptions: (1) both variants LIST and CATALOG being self-determinate are not frequently represented within genitive constructions (the seventh hypothesis) and (2) LIST in

77
genitive constructions favors less written registers (newspaper, fiction or academic) (the eighth hypothesis).

3. The method
To model the joint impact of the features presented above on choice between LIST and CATALOG, I draw on logistic regression analysis. Multiple logistic regression performed in R quantifies the effect of individual explanatory factors on a binary dependent variable, such as the effect of repeatable characteristics.

The mechanism of multiple logistic regression also allows to measure the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it considers the calculation of the most suitable option in the quantitative way. By means of regression I seek to predict a probability of application of a concept LIST as a computational unit and to build a strong base to avoid inaccuracies in the usage of terminology at further researches on semantics of ordering and enumerative units. The Performance of this task is conducted by comparing the predictors of concepts LIST and CATALOG.

3.1. The response variable
After having retrieved 1600 samples from the COCA, we annotated them for the variables below.

The response variable is encoded automatically with two values, CATALOG and LIST, with LIST as a success category.

3.2. The predictor register
Earlier we made the assumption that the CATALOG is used mainly in the strict professional and academic environment and is drawn towards a specification of the field of subjects covered by its semantics whereas the LIST has more common semantics and is respectively more open for a wide application.

In our work we considered distributions of these words across four registers of the language (academic, spoken, fiction, newspaper) on the data of the COCA.

Further we calculated a percentage ratio of the distributed cases in Excel according to 4 registers to the sum of all cases for the purpose of creation of more independent picture.

Then we allocated for the analysis 1600 occurrences in general and distributed 400 items in each register. On the basis of Table 1 containing a percentage ratio of words LIST and CATALOG we allocated their quantitative ratio.

Table 1

<table>
<thead>
<tr>
<th></th>
<th>Academic</th>
<th>Spoken</th>
<th>Fiction</th>
<th>Newspaper</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATALOG</td>
<td>893</td>
<td>117</td>
<td>362</td>
<td>423</td>
</tr>
<tr>
<td></td>
<td>7%</td>
<td>1%</td>
<td>5%</td>
<td>2%</td>
</tr>
<tr>
<td>LIST</td>
<td>11926</td>
<td>11181</td>
<td>6628</td>
<td>17812</td>
</tr>
<tr>
<td></td>
<td>93%</td>
<td>99%</td>
<td>95%</td>
<td>98%</td>
</tr>
<tr>
<td>Total</td>
<td>12819</td>
<td>11298</td>
<td>6990</td>
<td>18235</td>
</tr>
</tbody>
</table>
Based on the data of Table 1 I draw a conclusion that the frequency rate of the word LIST considerably exceeds one of CATALOG. This fact corroborates our hypothesis about the scope of the word CATALOG in non-technical registers which is much narrower than the LIST one (2).

The additional conclusion for further research on semantics of CATALOG and LIST is about rather low realization regarding other registers of both words in the fiction register that means these concepts are not often represented in the non-formal context.

Table 2 shows that we observe CATALOG mainly in specialized, professional and academic contexts whereas LIST is inherent for both formal and spoken contextual environment. In spite of the fact that the percent of LIST samples at spoken register is rather high (11 181 occurrences), the newspaper register (17 812 occurrences) still remains the most appropriate for the word LIST. Hence the hypothesis that LIST is inherent for contextual environment of a spoken language partly came true.

The purpose of the predictor register is to show whether non-technical registers favor LIST in any layout, that is when REGISTER, SCHEDULE, NOMENCLATURE, ROLL, INVENTORY are left out of the picture.

3.3. The predictor QUANTIFIED POSSESSIVE MEANING

The semantic feature of possessive meaning that we analyze in our case study is based firstly on narrowing and uniformity effects within the utterances with the notions LIST and CATALOG. Barker points out that quantification property, termed possessive existential import, is intimately related to the notion of narrowing [Barker 1995], i.e. narrowing has implications for compositional analysis of the meaning. We apply the proposed semantics to the issue of definiteness extent of LIST and CATALOG. Even though both words can be used to illustrate the narrowing effect, LIST seems to exhibit more versatility.

Considering LIST and CATALOG we notice that they are not only automatically narrowed by possessives, but become a kind of non-symmetric quantified units. Thus particular quantified possessors present LIST and CATALOG as the phenomenon of quantification. Examples include:

(4) Not every of his list is as good as that one.

(5) Our list—a scrolling queue of names that lined his left forearm—was always changing.

(6) Most of my catalogs are better than in previous years.

Additionally we notice that presupposition is the pragmatic mechanism that enforces narrowing [Beaver 2001]. It might be suggested that this mechanism inserts a presupposition of one part of a sentence’s meaning in another. Thus we can quantify narrowing by limiting to range only pre-possessors.

Hence:

1. Possessives are capable of narrowing the domain of quantification;
2. Possessive’s narrowing ability is related to the common feature of the possessor quantification;
3. Narrowing can be accomplished by a quantification of the cases of pre-possession.

The presence of quantified possessive meaning is simply encoded as “yes” and the absence as “no”. In the dataset of 1600 observations, there are 281 cases of poss. meaning = yes and 1319 cases of poss. meaning = no

3.4. The predictor PLURAL MODIFICATION

The probability of the use of LIST and CATALOG in plural form with a possessive pronoun increases in the presence of a negative estimated component in the semantics of the closest context.

(13) Desire to be on these lists might unintentionally be contributing to not just the growth in AP test-taking revealed.

(14) Visualizing context and his knowledge of images in the earlier literature isn’t evident in his catalogs.

The main condition of using LIST in the form of plural is the contextual specification of noun meaning which can be carried out in two ways: as sampling in CATALOG case and unitarization of a lexical meaning of a noun in the LIST case.

(15) News and World Reports to determine lists of "Best High Schools" of strongly integrates AP and International Baccalaureate.

(16) The 2005 workshop FRBR in 21st Century Catalogs (FRBR Workshop) is the first known venue.

We observe the similar division in the combination with numerals and other quantitative determinants, and also in contexts with the indirect instruc-
tion on plurality of the considered concepts of LIST and CATALOG.

(17) There were 21982 patients in the USA on waiting several lists for various organs but only 4248 cadaveric organ donors.

(18) Curators and educators as well take the information provided by these five catalogs as factual.

In these cases they gain ability to transfer quantitative feature. It is possible to call this process quantification of LIST and CATALOG and in this regard to distinguish two ways of quantification: discrete and unitary.

There are 6 cases of plural modification and 1594 cases in singular form.

3.5. The predictor PRE-MODIFICATION

«Pre-modifiers are condensed structures. They use fewer words than post-modifiers to convey roughly the same information. Most adjectival and participial pre-modifiers can be rephrased as a longer, post-modifying relative clause...» [Pearson 2002]

LIST being less formal concept is more drawn towards pre-modification. There are certain types of contexts in which lists are rather consistently and regularly used in PRE.MOD utterances.

There are four major structural types of pre-modification in English:
- general adjective: (19) official list, (20) new catalog;
- -ed participial modifier: (21) restricted list, (22) established catalog;
- -ing participial modifier: (23) waiting list, (24) publishing catalog;
- noun: (25) staff list, (26) sales catalog

In addition, determiners, genitives, and numerals precede the head and modifiers, and help to specify the reference of noun phrases.

There are 868 cases of pre.mod = yes and 732 cases of pre.mod = no.

3.6. The predictor POST-MODIFICATION

We allocate the predictor POST.MOD which stands for post-modification. In this respect there are three values of post-modifiers: (1) prepositional phrases, (2) finite relative clauses, (3) ed-clauses and appositive noun phrases. These values are rather common for written language [Biber 1999]. All instances of phrases that consist of a preposition followed by another word, phrase or clause functioning as a prepositional complement are defined as prepositional phrases.

(27) Data for our catalog;
(28) The list of guests;

Finite relative clauses have three types: which/who-clauses, that-clauses and non-finite clauses.

(29) I have a fantastic catalog which nobody has seen before;

(30) The librarian showed the catalog that contained all the milestones of the time;
(31) It was a list to revive the ancient traditional dish;

Finally ed-clauses and appositive noun phrases seek to add new features to the noun and make the utterance more interesting.

(32) Catalog presented in the Chapters below;
(33) Synonyms, list of frequencies.

They are used in order to make a sentence more evident, to reflect a more particular component of the noun or to rename a noun beside it.

The presence of POST.MOD indicates the high level of internal componential structure of words, as well as the capacity of words to interact with other words and phrases, acting as a measure of something or explanatory tool based on the principle of similarity.

The purpose of the variable POST.MOD is to test, firstly, that both variants are post-modified, secondly, that LIST favors a more frequent post-modification.

There are 890 cases of post.mod= yes, 710 cases of post.mod= no.

3.7. The predictor GENITIVE MEANING

We split up genitive and possessive meanings due to the fact that the first has functions besides the possessive ones, such as:
- (1) Indication of the whole/part relations: (34) three pages of the list, (35) cover of the catalog;
- (2) Allocation of qualitative features: (36) A strict mind of catalog, (37) We discuss an exceptional value of lists;
- (3) Designation of the subject of action: (38) A compilation of lists, (39) In view of the growing popularity among young generation was planned a reissue of this catalog;
- (4) Reference to the object of action: (40) A catalog’s ordering, (41) A list’s enumeration;
- (5) Quality, condition or action holder denotation: (42) Relevance of the catalog, (43) Length of the list;
- (6) Explanation of a word to which it refers: (44) A set of the list, (45) A text of the catalog;
- (7) Introduction of quantitative restrictions: (46) In the modern world occur a lot of lists, (47) Most of the catalogs are built on the alphabetical principle.

Another argument in favor of genitive meaning allocation apart from the possessive one is submitted by John Lyons [Lyons 1968: 394-395] and Charles Fillmore [Fillmore 1968: 49-50]. They consider that a derivation cannot account for the wide range of semantic readings available in many’s genitive constructions.

‘S-genitives and of-genitives are considered as two roughly equivalent ways of saying the same
thing. [Labov 1972] Therefore we include of-
genitive and s-genitive constructions in the analysis
within the meanings outlined above. There are 371
cases of genitive meaning = yes and 1229 cases of
genitive meaning = no in the corpus.

4. Concluding remarks

Elaboration of the preparatory stage of the anal-
ysis in the logistic regression acts an important role
in producing quality results. Set parameters deter-
mine the course of the study case. All advanced
hypotheses delimit the scope and do for verification
models.

Explanatory notes

2 Ibid.
3 The examples are taken from the site: Da-
vies M. The Corpus of Contemporary American
English (COCA): 425 million words, 1990, URL:
www.americancorpus.org.

4 a case of chaotic enumeration, where apparently
there is no connection between different elements
of the list. Nevertheless, it is worth dusting a distinction
off between conjunctive enumeration and disjunctive
one. A conjunctive enumeration assembles also
different things that makes the set to be coherent, since
the subject itself consider them in an identical con-
text; on the contrary, the disjunctive enumeration ex-
presses a fragmentation, a species of schizophrenia of
the subject that conceives a sequence of ludicrous
impressions without any utility feature.

References

Barker Ch. Possessive Descriptions. Stanford:

Beaver D. Presupposition and Assertion in Dy-
namic Semantics. Stanford: CSLI Publications,

Biber D., Conrad S., Leech G. Longman Gram-
1204 p.

Charles W. G. Contextual Correlates of Mean-
P. 505–524.

Collins – Collins English Dictionary / Ed. by

Collins – The New Collins Thesaurus / Ed.

Church K., Gale W., Hanks P., Hindle D.,
Moon R. Lexical Substitutability // B. T. S. Atkins
and A. Zampolli. Computational Approaches to the

Davies M. The Corpus of Contemporary Amer-
URL: http://www.americancorpus.org (дата обра-
щения: 11.06.2016).

Eco U. The infinity of lists. Pub: New York: Rizz-

Farlex – Free Farlex Online Thesaurus of syno-
onyms, antonyms, and related words. URL: http://
www.thefreedictionary.com/list (дата обращения:
09.06.2016).

Fillmore Ch. “The Case for Case”, In Universals
in Linguistic Theory / Ed. by Emmon Bach and

Gilquin G. Causative ‘get’ and ‘have’: So close,
№ 31(2). P. 125–148.

Glynn D. Mapping meaning. Towards a usage-
based methodology in Cognitive Semantics. (Un-
published doctoral dissertation). University of Leu-

Gries S. Th. A corpus-linguistic analysis of -ic
P. 65–108.

Gries S. Th. & Otani N. Behavioral profiles:
A corpus-based perspective on synonymy and an-

Halliday M. A. K. Spoken and written modes of
meaning / ed. by Rosalind Horowitz and S. Jay
Samuels. Comprehending Oral and Written Lan-

Hunston S. Corpora in applied linguistics. Cam-

Labov W. The recent history of some dialect
markers on the island of Martha’s Vineyard Mass /
ed. by L. M. Davis. Studies in Linguistics in Honor
with Malcah Yaeger & Richard Steiner. A Quantita-
tive Study of Sound Change in Progress. Philadel-

Levshina N., Geeraerts D., Speelman D. Dutch
causative constructions with doen and laten: Quantifi-
cation of meaning and meaning of quantification / ed.
by D. Glynn & J. Robinson. Corpus Methods for Se-
manitics: Quantitative studies in polysemy and synon-

Longman – Longman Dictionary of Contempo-
rary English. Fifth edition / Ed.: Harlow, Pearson,

Lyons J. Introduction to Theoretical Linguistics.

Miller G. A., Walter G. Ch. Contextual correlates
of semantic similarity. Language and Cognitive Pro-

Peters S. The Semantics of Possessives. Stanford
University Press, 2013. 64 p.

Phochoaroensil S. A. Corpus-Based Study of
English Synonyms. International Journal of Arts and

Webster – Random house Webster’s college dic-
1567 p.
In this case study we present two near-synonymous concepts which reflect the human outlook. The aim of this work is to define the most appropriate one for its usage in the further investigations on semantics of ordering and enumerative units. Being unconscious and objective, these units exist irrespectively of the observer and can be defined by distinctions and oppositions. From this point of view we get LIST and CATALOG as near-synonymous words seeking to identify the factors that determine the choice of a concept by objective and semantic criteria. For this purpose we elaborate an outline of multiple logistic regression analysis. The focus of investigation is on the identification of the contents and dynamics of the concepts verbalized by LIST and CATALOG across (1) four registers of the language (academic, spoken, fiction, newspaper), (2) semantic feature of possessive meaning, (3) plural modification, (4) post-modification, (5) pre-modification and (6) genitive meaning.

Key words: computational linguistics; logistic regression; comparative analysis; semantics; near-synonymous concepts.