The Experimental Study of Terminology Collocations: Calculations and Experiments with Informants

Elena Yagunova
Saint-Petersburg State University
Saint-Petersburg, Russia
iagounova.elena@gmail.com

Anna Savina
Saint-Petersburg State University
Saint-Petersburg, Russia
anja.savina@gmail.com

Abstract

This paper presents an experimental solution of the problem of the nature of the terminology collocations and possibility of their ranging, which depends on the degree of coherence of these collocations. Within this paper the combination of two different approaches – calculation and experiments with informants – is proposed to the study of the terminology collocations. The proposed approach is particularly relevant for those scientific areas, where still there isn't precise terminology.

1 Intoduction

Our research is devoted to solving one of the most important problems of collocation study: about the nature of scientific (terminology primarily) collocations and their possible classification. The report presents the result of the first stage of work within the overall project on this topic. We understand a **collocation** as a non-random combination of two or more lexical items that characterizes the language as a whole (texts of any kind) or a certain text type (or even (sub)sample of texts). In our research of language and speech we go from the text **realization**, from the available material. The material dictates the choice of certain theoretical positions and classification principles. Such research may be conducted only by using statistical measures to evaluate the degree of the non-randomness of the sequence of words. It's obvious that the list of combinations isn't completely homogeneous and requires a subsequent classification and some theoretical interpretation (Pivovarova and Yagunova, 2010; Khokhlova, 2011; Yagunova and Pivovarova, 2011).

Ample opportunities for understanding the nature of the collocations – within the lists received on the basis of statistical measures – are given by the reference to experiments with the native speaker informants. The purpose of the report is the demonstration of such kind of capabilities.

2 Material and Methods

We want to illustrate the suggested methodology based on an example of a monothematic collection of conference materials "Corpus Linguistics" for 2004-2008[1]. Volume of the collection is about 220,000 "tokens" – word usages and punctuation marks. Corpus Linguistics (especially in Russia) is the scientific area, where still there isn't precise terminology.

We used two statistical measures for bigram extraction: MI and T-score (Evert, 2004; Manning and Schütze, 1999; Stubbs, 1995). MI allows to extract terminological combinations, T-score highlights scientific clichés and those terminological combinations that characterize all texts from the collection (or most of those texts) (Pivovarova and Yagunova, 2010). A.Savina has made the program, which is very convenient for research purposes and allows to select lists of bigrams with a nuclear word on the basis of those statistical measures.

We have considered two ways to get lists of interesting (for us) collocations:

- for all collocations with a maximum value of these measures;
- for collocations with interesting (for us) nuclear word.

[1] For comparison, we used a collection of news texts lenta.ru in 2009. Thus, we tried to research collocation, describing the text of a certain functional style (type, genre).

Based on the collection of scientific texts we first of all had received lists of bigrams for each of the measures of association (MI and T-score) and sorted them by descending value of these measures. Then from each list were selected from 25 bigrams with the highest values of the measure.

After that for each nuclear word (“corpus”, “word”) we also consider sublists of 25 bigrams for each of the measures.

Later on the combined listings of the 50 randomized collocations are the input data for the two types of experiments with informants.

Such combined lists allowed us to estimate the degree of connectivity for terminological combinations, which was allocated on the basis of measures MI, and then compared with scientific clichés, function words (and other combinations), allocated on the basis of the measures T-score.

For this monothematic collection terms, that are common to all texts of the collection, were distinguished on the basis of both measures.

Thus, we obtained two lists of bigrams – with nuclear words and without.

As it has been already mentioned, we conducted two types of experiments with each of the lists:

- experiment 1 – the classification of bigrams;
- experiment 2 – the scaling of bigrams.

Experiment 1: 25 informants offered informants a questionnaire in which they were required to determine which of the three classes – “right”, “predictable” and “others” – applies to each combination of the proposed list.

Experiment 2: 22 informants were given the task to evaluate the degree of connectedness between words – for the same lists – at a scale of 0 to 5, where “0” corresponds to the minimum, and “5” – the maximum degree of connectivity from the perspective of informants.

In the instructions we said to the informants about the domain specificity: “You see the combinations of words (bigrams) from the specialized (linguistic conference) texts, selected on the basis of statistical criteria. ...”

3 Results. Conclusions

We have obtained the classification of bigrams in a given set of classes for a holistic list (without specifying the nuclear word) based on the results of Experiment 1. Each class is divided into the core and the periphery. Collocation was considered as core, if more than 65% of informants referred it to this class and the peripheral – if the number of informants ranged from 33% to 65% (amount of these classes in table 1)[2].

Table 1. Bigram classes according to Experiment 1

bigrams	“right”		“predictable”	
classes	core	periphery	core	periphery
without nuclear words	12	12	6	27
with nuclear words	9	6	10	15

Table 2. The results of Experiment 1 (bigrams without nuclear words)

core of “right”	core of “predictable”	core of “others”
математической лингвистики [mathematical linguistics]	в качестве [as]	но и [but also]
художественной литературы [fiction]	за счет [due to]	так и [and]
русского языка [the Russian language]	(в) свою очередь [(in) turn of]	что в [that in]
корпусная лингвистика [corpus linguistics]	(в) том числе [including]	
имена собственные [proper names]	на основе [on the basis of]	
словарной статьи [vocabulary entry]	(с) точки зрения [(in) terms of]	
машинного перевода [machine translation]		
корпусной лингвистике [corpus linguistics]		
корпусной лингвистики [corpus linguistics]		
речевой деятельности		

[2] Part of peripheral collocations was attributed to the intersection of classes (unless this requirement observed with respect to two classes).

Continuation of Table 2. The results of Experiment 1 (bigrams without nuclear words)

[speech perception and speech production]		
XX века [XX century]		
корпуса текстов [text corpora]		

The term *"контекстной предсказуемости [context predictability]"* is a very clear example of differentiation between statistical (maximum of MI-score) and informant-used approaches: our informants attributed this term to the intersection between "right' and "predictable". What does it mean? Maybe this term is not wide-used, or it belongs to the other domain, but degree of intuitive connectedness and wholeness of the bigram is more important then statistical one for many terminology classification tasks.

We illustrate the capabilities of our method of analysis on a sample list of bigrams (for word forms bigrams) with the nuclear word "corpus" and "word". 42% of collocations were related to nuclear bigrams:

- the core of "right" contains scientific terms (or their components);
- the core of "predictable" contains mainly compound words;
- the core of "others" – a combination that is difficult to interpret.

Experiment 2 has allowed us to establish the degree of connectedness between the components of bigrams and define flexible boundaries between classes.

The data of Experiment 2 verifies those hypotheses on the classification that has been received as the results of Experiment 1. The list of bigrams, the connection of which is estimated by a group of informants is not less than 4 points (average of the group), fully consistent with the core of "right" (according to Experiment 1). These bigrams are allocated on the basis of MI (and sometimes T-score). Those bigrams whose connection is more than 2,8 – the core of the "predictable" (according to Experiment 1).

The core of "others" (the connection is less than 2,8), these are high-mix combinations, which could not be cut off by the correction factor to the extent of T-score.

We have obtained similar results for bigrams with the nuclear word – "corpus" or "word" – as in Experiment 1: the group "core of "right" was just terminological combinations (Table 3).

Table 3. The results of Experiment 1 (bigrams with the nuclear word)

core of "right"	core of "predictable"
аннотированный корпус [annotated corpus]	корпус является [corpus is]
национальный корпус [national corpus]	корпус содержит [corpus contains]
параллельный корпус [parallel corpus]	корпус представляет [corpus represents]
международный корпус [international corpus]	корпус позволяет [corpus allows]
представительный корпус [representative corpus]	данный корпус [this corpus]
размеченный корпус [labeled corpus]	большой корпус [large corpus]
электронный корпус [electronic corpus]	второе слово [second word]
служебное слово [function word]	первое слово [first word]
составное слово [compositum]	данное слово [this word]
	слово встретилось [word is used]

The data analysis of Experiment 2, at first, confirmed the results of Experiment 1, i.e. all bigrams of the list "core of "right" had a value of the connection at least 4. Secondly, Experiment 2 expanded our list of the most associated bigrams by terminological combinations *"главное слово" [main word], "зависимое слово" [depended word], "отдельное слово" [separate word]*, which in experiment 1 were in the intersection of groups of "right" and "predictable" bigrams.

The results of Experiments 1 and 2 allow us to install additional scales, based not only on the values of statistical measures, but also of the feeling the degree of connectivity (by native speakers), which can become explicit during the experiments.

Thus, we propose an experimental approach, which combines the methods of computing experiment, and the experiments with informants. Terminological combinations are the most connected from the viewpoint of the experiment, even when compared with such units as Multiword function words (such as, *в качестве [as], в частности [in particular], за счет [due to]* and etc).

Multiword terminology gets explicit hierarchy in terms of the degree of connectivity within their own class. For example, from the perspective of our informants (students after corpus linguistics lectures), there are several levels of this kind of connectedness (in descending order), see Table 4 with levels of connectedness from two different perspectives.

Table 4. Examples of levels of connectedness

From the perspective of our informants	From the perspective of MI-score
художественной литературы [fiction]	контекстной предсказуемости [context predictability]
математической лингвистики [mathematical linguistics], корпусной лингвистике [corpus linguistics], имен собственных [proper names], корпусная лингвистика [corpus linguistics], имена собственные [proper names], машинного перевода [machine translation], корпусной лингвистики [corpus linguistics], корпуса текстов [text corpora]	речевой деятельности [speech perception and speech production], художественной литературы [fiction], имен собственных [proper names], корпусная лингвистика [corpus linguistics], имена собственные [proper names]
контекстной предсказуемости [context predictability], речевой деятельности [speech perception and speech production], русского языка [the Russian language]	математической лингвистики [mathematical linguistics], словарной статьи [vocabulary entry]
предметной области [(knowledge) domain]	предметной области [(knowledge) domain]
словарной статьи [vocabulary entry]	машинного перевода [machine translation],

We don’t pretend to give an exhaustive description of terminology (for example, the terminology of corpus linguistics). However, the results allow us to take a fresh look at the application of terminology. It is particularly relevant to some new scientific paradigms or to interdisciplinary areas, where there is a process of formation of terminology. In our opinion, the proposed approach has a great future in the study of terminology and predicting the potential of different terminology variants.

Our goal is not only in terminology extracting. Set of terms will not be homogeneous. The advantages of this approach – combined the corpus based and informant-used methods – are in classifying of multiword terminology. This classification must have both statistical and human (informants and/or experts) basis.

We plan to compare the application of terminology in some different domains. We have got a monothematic scientific collection “Hydrogeology” (more then 200 000 “tokens”). Hydrogeology is semantically less closely related to the researchers’ field of computational linguistics. We suppose also that hydrogeology has more precise terminology then corpus linguistics. The next step is the comparison of multiword hydrogeology terminology with the terminology of corpus linguistics.

References

Stefan Evert. 2004. *The Statistics of Word Cooccurences: Word Pairs and Collocations*. PhD thesis. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

Chris Manning, Hinrich Schütze. 1999. Collocations. *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA.

Michael Stubbs. 1995. Collocations and semantic profiles: on the case of the trouble with quantitative studies. *Functions of language* 2(1): 23-55, Benjamins.

Lidia Pivovarova, Elena Yagunova. 2010. Extraction and Classification of Terminology Collocations of the Material of Scientific Texts (preliminary observations). *Materials of II International Symposium “Terminology and Knowledge”*. Moscow.

Maria Khokhlova. 2011. The Research of Lexical-semantic Compatibility in Russian Language with Statistical Measures (on basis of corpora). AKD Saint-Petersburg.

Elena Yagunova, Lidia Pivovarova. 2011. From Collocations To Constructions. *Russian Language: Structural and Lexical-semantic approaches*. Saint-Petersburg.