# New Method of Automated Terminology Extraction: Case Study of Russian-language Textbooks

Sergei Monakhov, Vladimir Turchanenko,
Ekaterina Fedyukova, Dmitry Cherdakov

FTC 2021

The reported study was funded by RFBR, project number 19-29-14032 mk

# Traditional statistical approach

Two corpora are employed:

- the target corpus from which the terminology is extracted,
- the reference corpus, normally the national corpus of the relevant language.

# Disadvantages of traditional approach:

- it does not extract low-frequency terms,
- it results in a list of terms ordered by frequency, but does not provide a comprehensive image of the terminological system these terms belong to,
- the resulting lists of term-candidates are plagued with infrequent lexical units which are for some reason widely represented in the target corpus under consideration.

# Modern trends

- Context holds crucial information about the semantics of the word.

- Measure of specific lexemes' semantic proximity is calculated as the probability of their co-occurrence within a certain distance of each other.

# Research corpus of textbooks in Russian

- 212 items in 21 subjects; 14,370,000 words.

- By discipline: **algebra**—18 textbooks; **astronomy**—2 textbooks; **biology**—21 textbooks; **chemistry**—13 textbooks; **computer science**—6 textbooks; **crafts**—4 textbooks; **fine arts**—8 textbooks; **geography**—8 textbooks; **geometry**—8 textbooks; **law**—2 text- books; **literature**—36 textbooks; **mathematical analysis**—14 textbooks; **mathematics**—10 textbooks; **music**—4 textbooks; **natural science**—2 textbooks; **physical education**—7 textbooks; **physics**—15 textbooks; **Russian**—4 textbooks; **social studies**— 12 textbooks; **world art culture**—2 textbooks; **world history and Russian history**—17 textbooks.

# Automatic Terminology Extraction

- Reference corpus—Russian Web 2011 Sample (*ruTenTen11*), 900 millions of words.

- Different selection principles for individual words (*keywords* in Sketch Engine) and multi-word expressions (*terms* in Sketch Engine).

# Automatic Terminology Extraction: Keywords Score

**((Lt * 1,000,000 / Ct) +1) / ((Lr * 1,000,000 / Cr) +1)**,

where

**Lt** is the frequency of the lexical unit in the focus corpus,

**Ct** is the total number of tokens in the focus corpus,

**Lr** is the frequency of the lexical unit in the reference corpus,

**Cr** is the total number of tokens in the reference corpus.

# Automatic Terminology Extraction: Terms Score

$$14 + \log(2(|X \cap Y|) / (|X| + |Y|)),$$

where

**|X|** is the absolute frequency of the first element of the combination in the focus corpus,

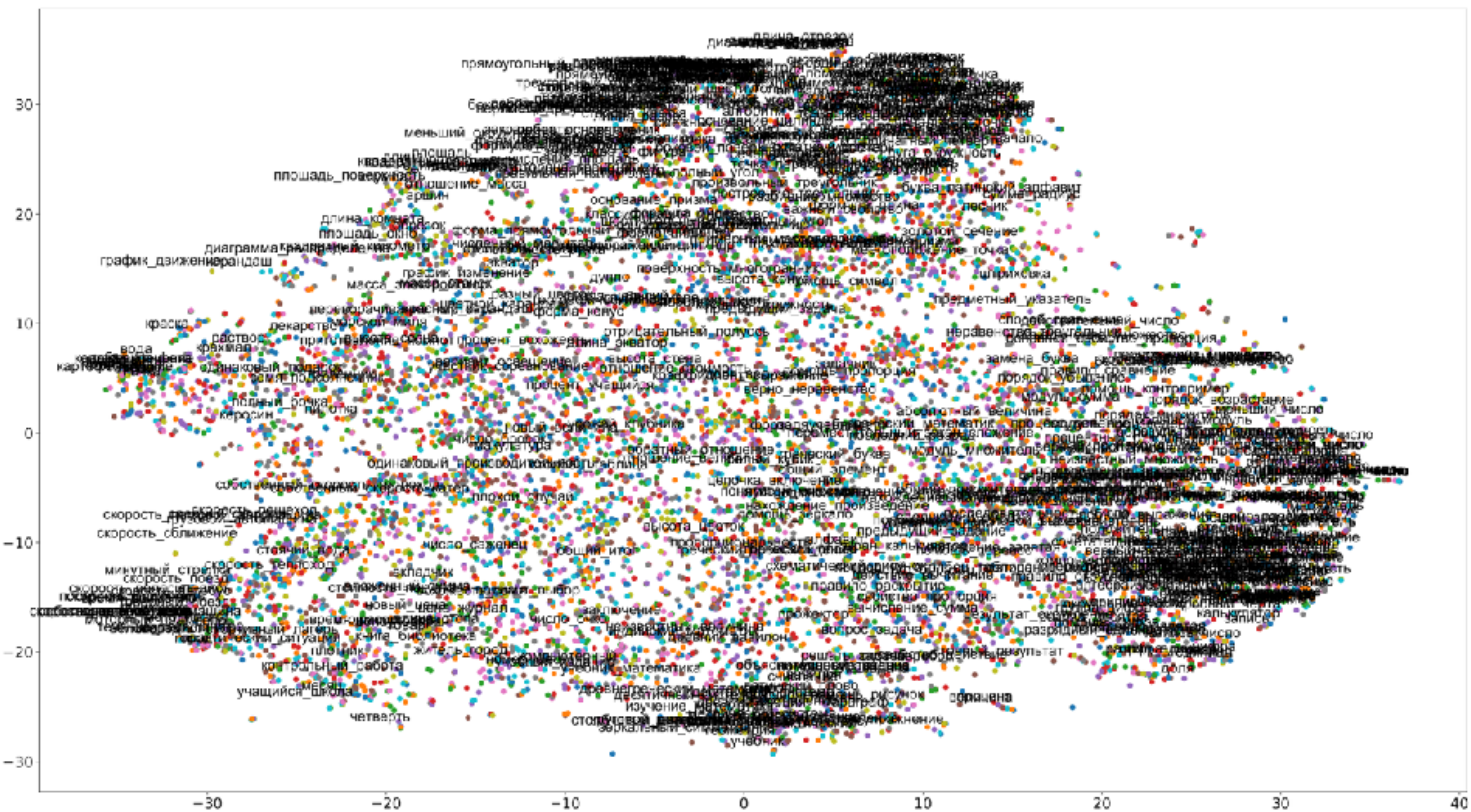
**|Y|** is the absolute frequency of the second element of the combination in the focus corpus,

**|X∩Y|** is the absolute frequency of the whole combination in the focus corpus.

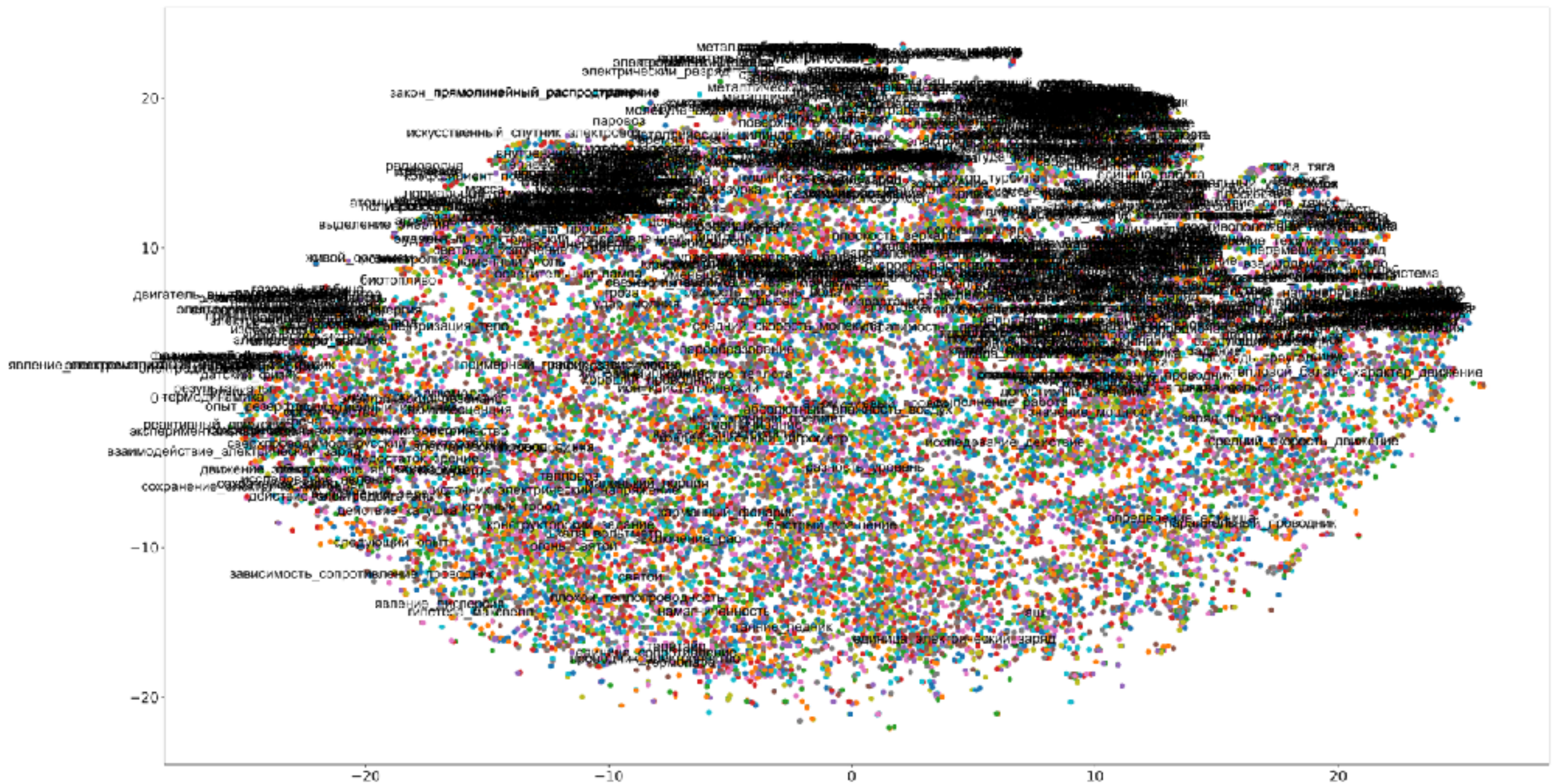
# Automatic Terminology Extraction: Word Embedding Models

- Target corpus was vectorised.
- Word-embedding models were trained for each area of knowledge.
- Maps reflecting the relative position of term-candidates in the obtained models were created and projected from a high-dimensional vector space into a two-dimensional plane using the t-distributed stochastic neighbour embedding (t-SNE) method.

# Automatic Terminology Extraction: t-SNE space visualisation



# Automatic Terminology Extraction: t-SNE space visualisation



# Automatic Terminology Extraction: Clustering

Clusters labelling was based on the following factors:

- the proportion of lexemes that occur within a cluster both as separate units and as part of word combinations (the hypothesis was that terminological clusters are characterised by a higher degree of repetition than non-terminological clusters),

- the proportion of multi-word combinations within a cluster (the hypothesis was that the number of multi-word units is higher in terminological clusters because automatic extraction of multi-word terms demonstrates a higher level of accuracy than extraction of individual terms).

Based on these factors, one common metric with a value for each cluster varying from 1 to 7200 was calculated. A value of 1 indicates that term-candidates in this cluster are highly likely to be terms. A value of 7200 indicates that term-candidates in this cluster are highly likely to be false terms.