USING OF THE STATISTICAL METHOD FOR AUTHORSHIP ATTRIBUTION OF THE TEXT

http://web.kpi.kharkov.ua/iks/

1st International Electronic Conference on Entropy and Its Applications, 3 - 21 November 2014

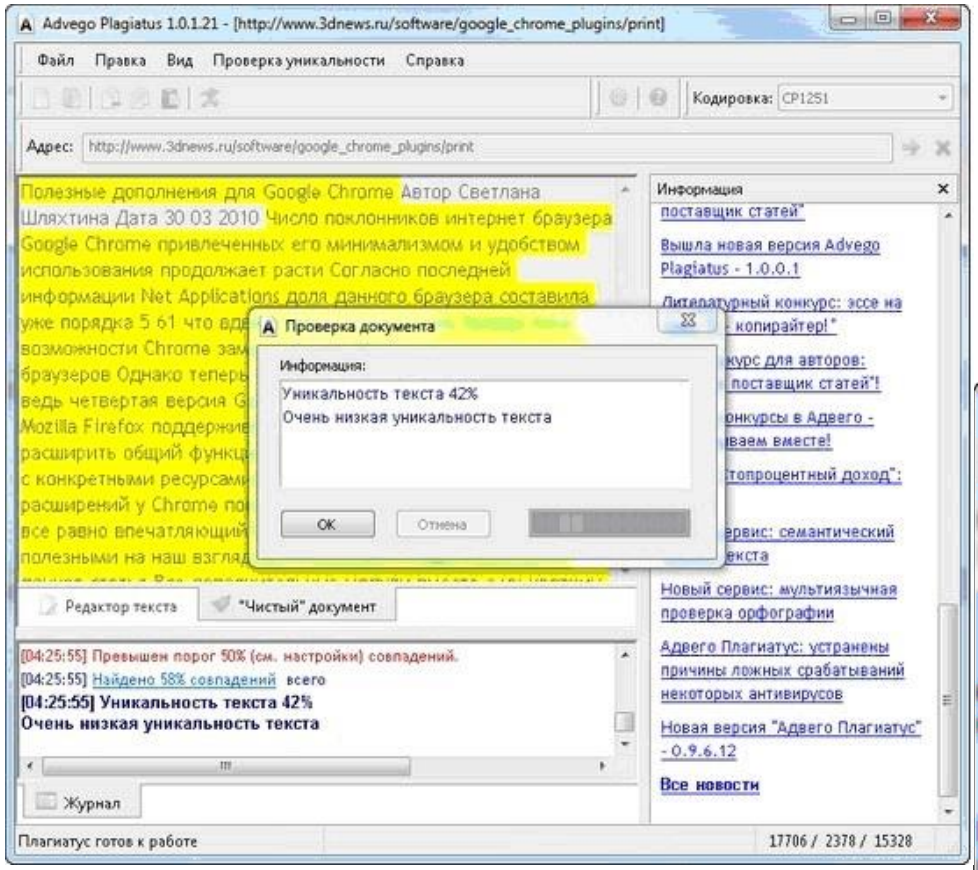
Purpose and objectives

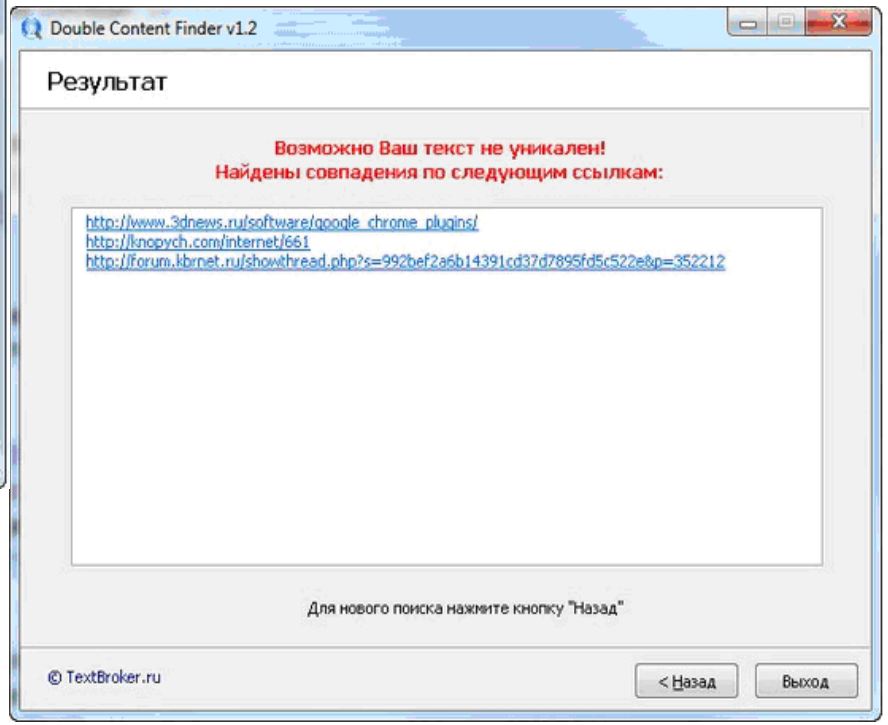
The aim of the thesis is to develop a system of attribution and methods for constructing pattern the author's style.

In accordance with the aim set the following tasks:

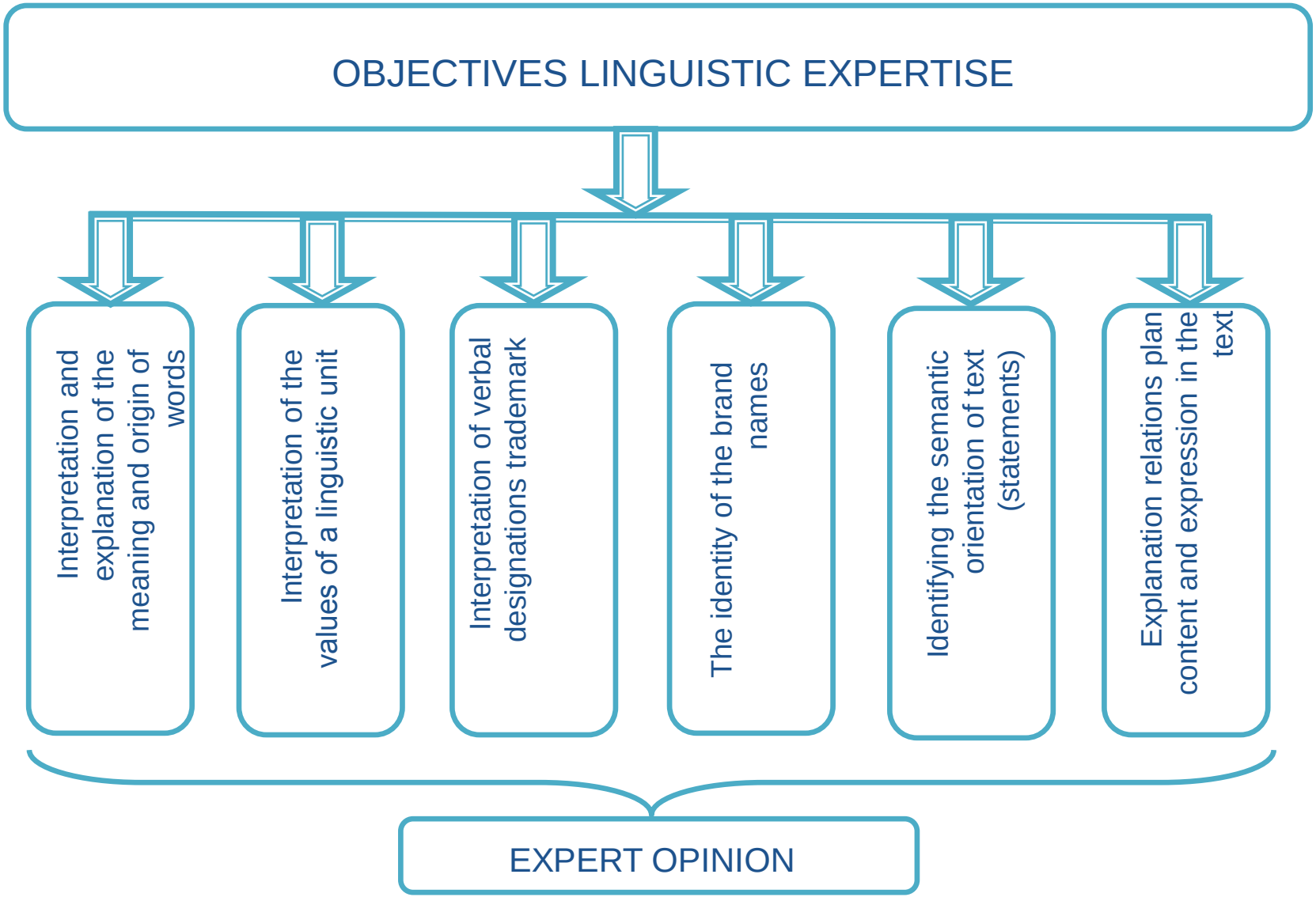
- determine the range of tasks that require linguistic expertise;
- consider the basic methods of automatic processing of texts for the construction pattern the author's style;
- to review existing systems that implement automated processing of texts;
- develop a linguistic model of the text, taking into account the statistical characteristics of the linguistic objects of author's text;
- develop the structure of the lexical database of language A. Pushkin;
- develop an algorithm for the formation of a lexical database language A. Pushkin;
- develop an algorithm for determining the adjacency of the texts;
- implement software implementation text recognition A. Pushkin.

Software systems

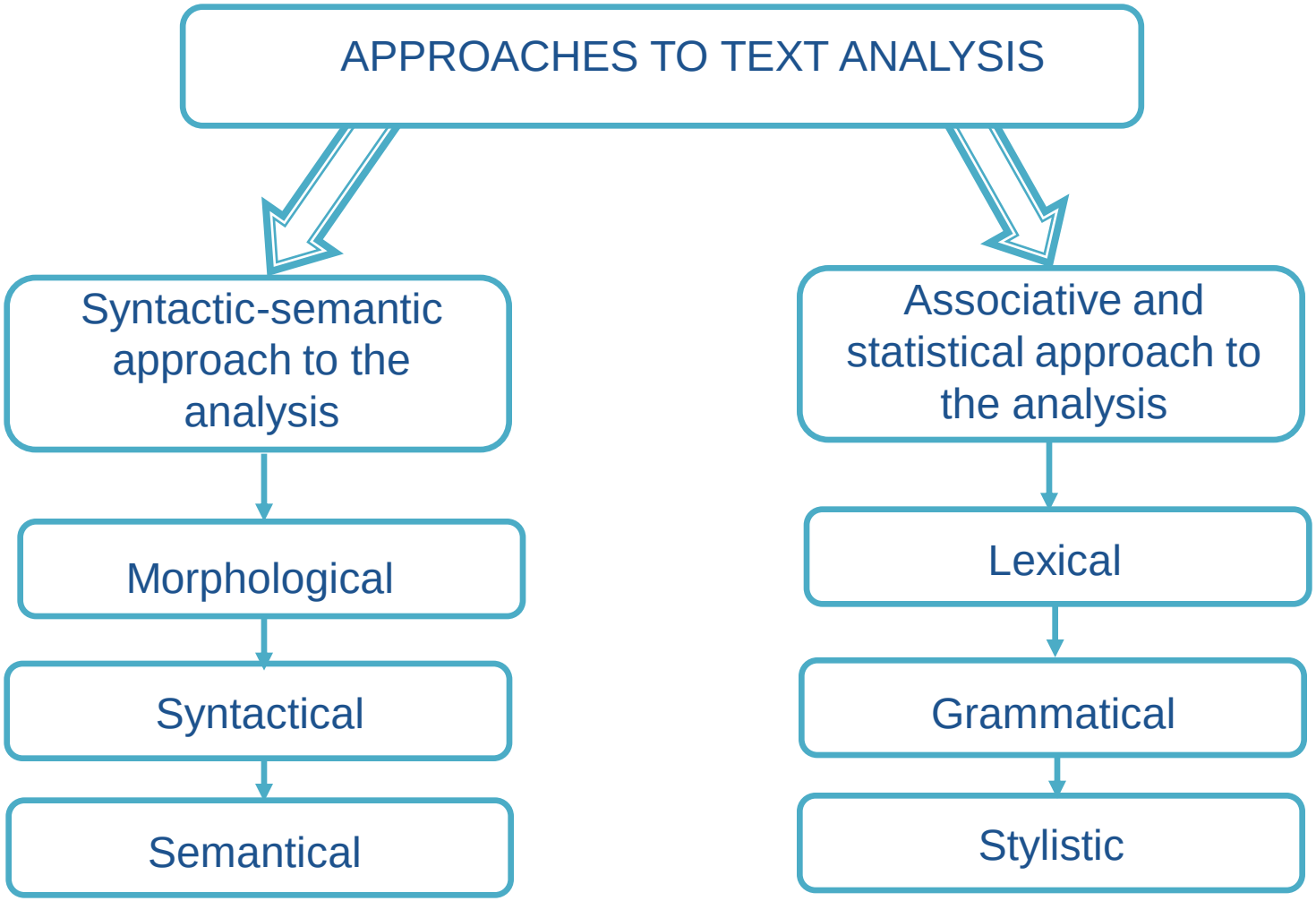




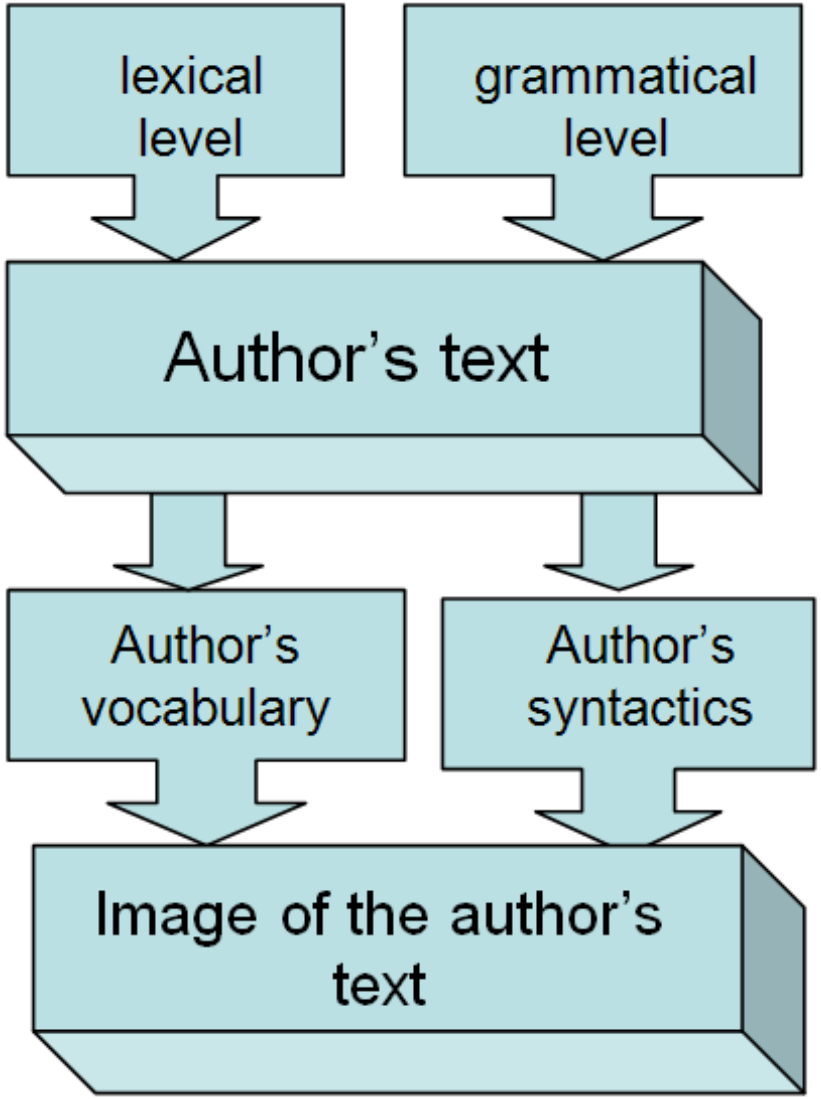
Goals linguistic expertise



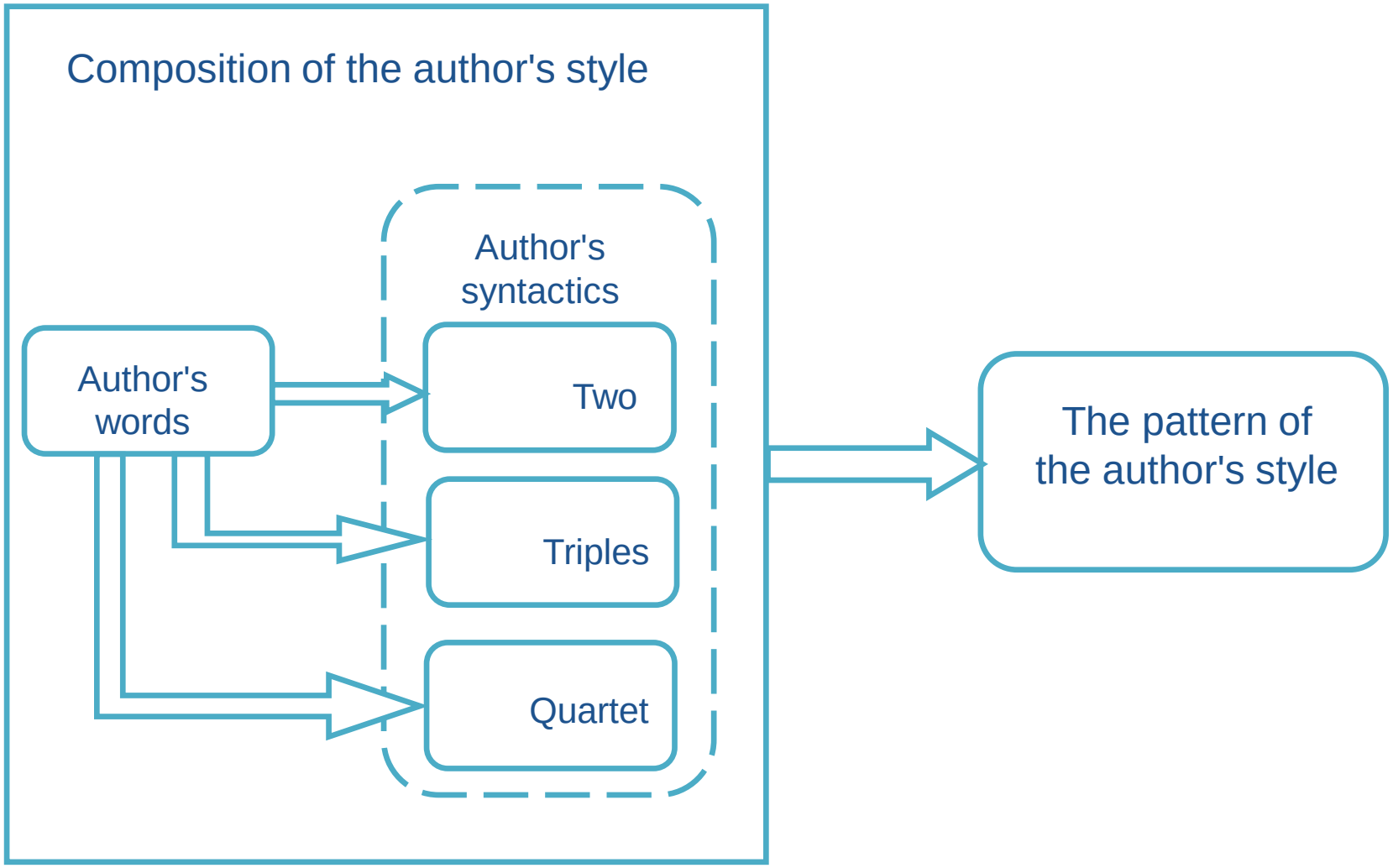
Approach in linguistic expertises



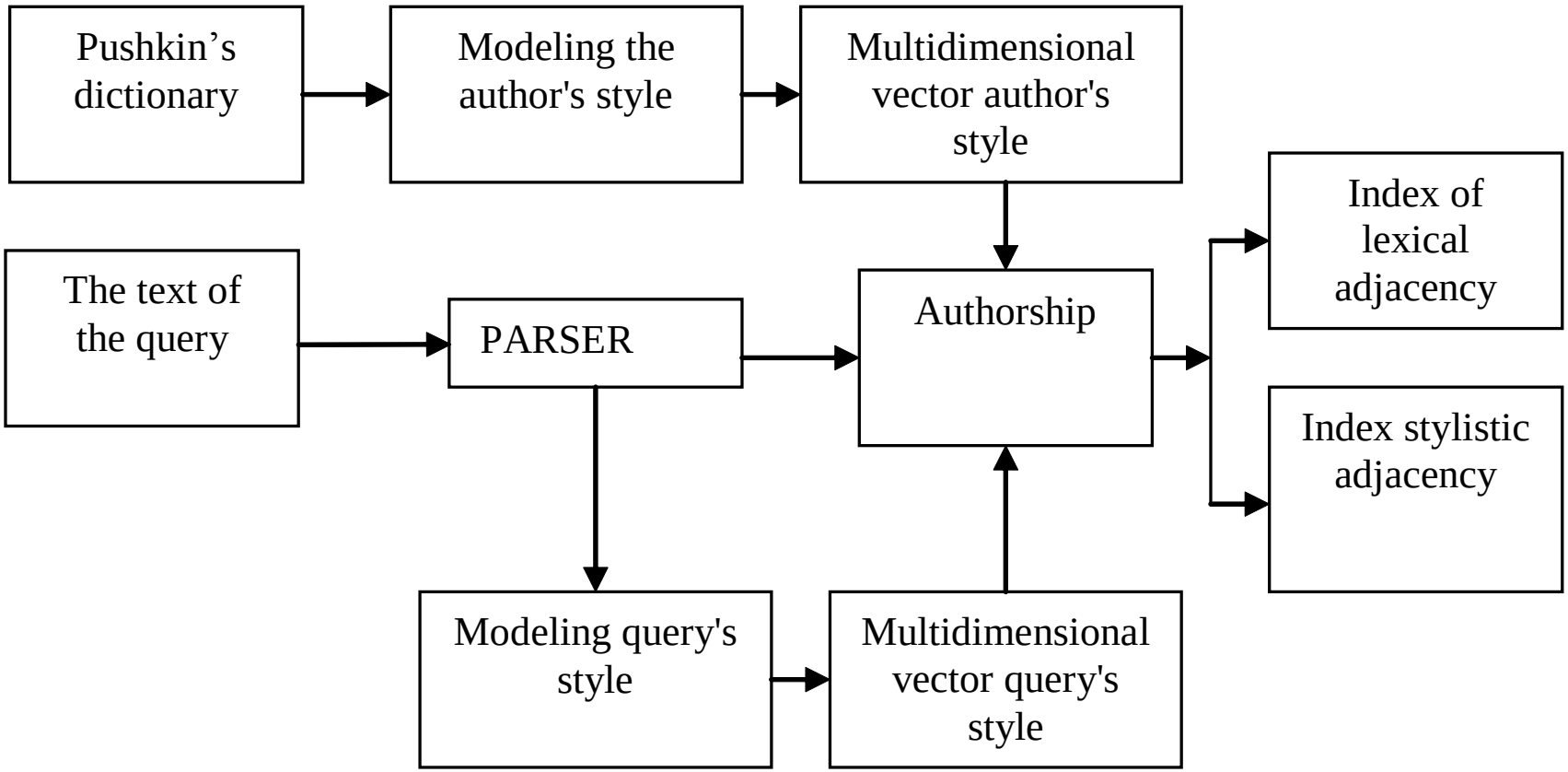
Scheme of a combination of approaches of lexical and grammatical levels



Modeling the author's style of Pushkin



Scheme of authorship



DB table of author's style

index_one

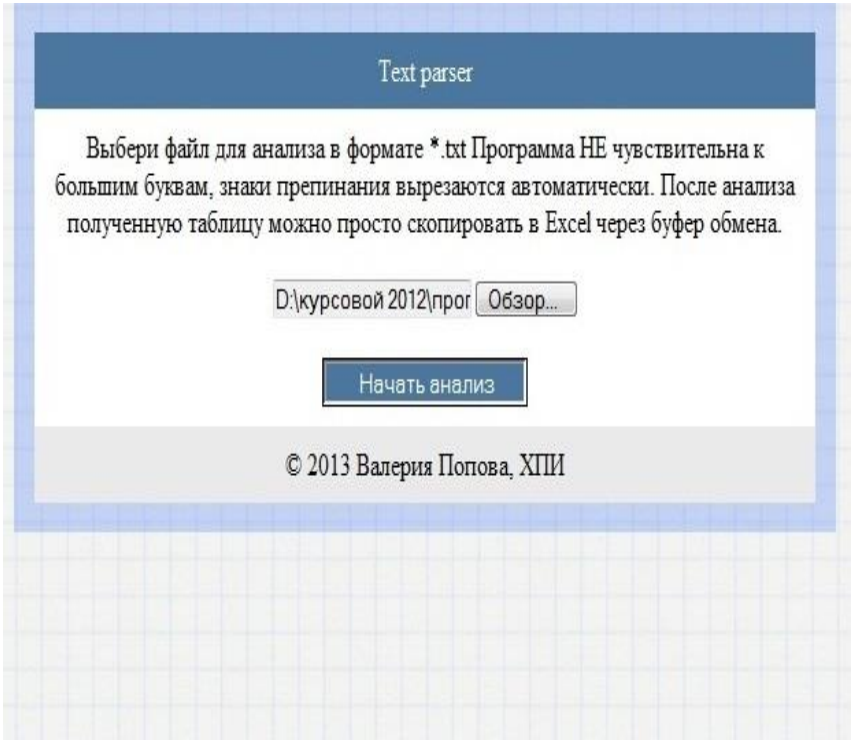
id	word	freq	Добавить поле
1	а	0,101	
2	береза	0,0667	
3	бы	0,033	
4	быть	0,0163	
5	в	0,0119	
6	весенний	0,0917	
7	весна	0,0833	
8	весь	0,0332	
9	ветер	0,0625	
10	вздыхать	0,0714	
11	взор	0,0833	
12	видать	0,1111	
13	внимать	0,1429	
14	восковой	0,1111	
15	встречать	0,0667	
16	вы	0,0564	
17	вылетать	0,1429	
18	глаз	0,0909	
19	глас	0,1429	
20	гостья	0,1111	
21	дорогой	0,1111	
22	душистый	0,125	
23	дуть	0,0667	
24	его	0,125	
25	еще	0,1111	
26	же	0,0366	
27	за	0,029	
28	замечать	0,0909	
29	зацветать	0,125	

index_two

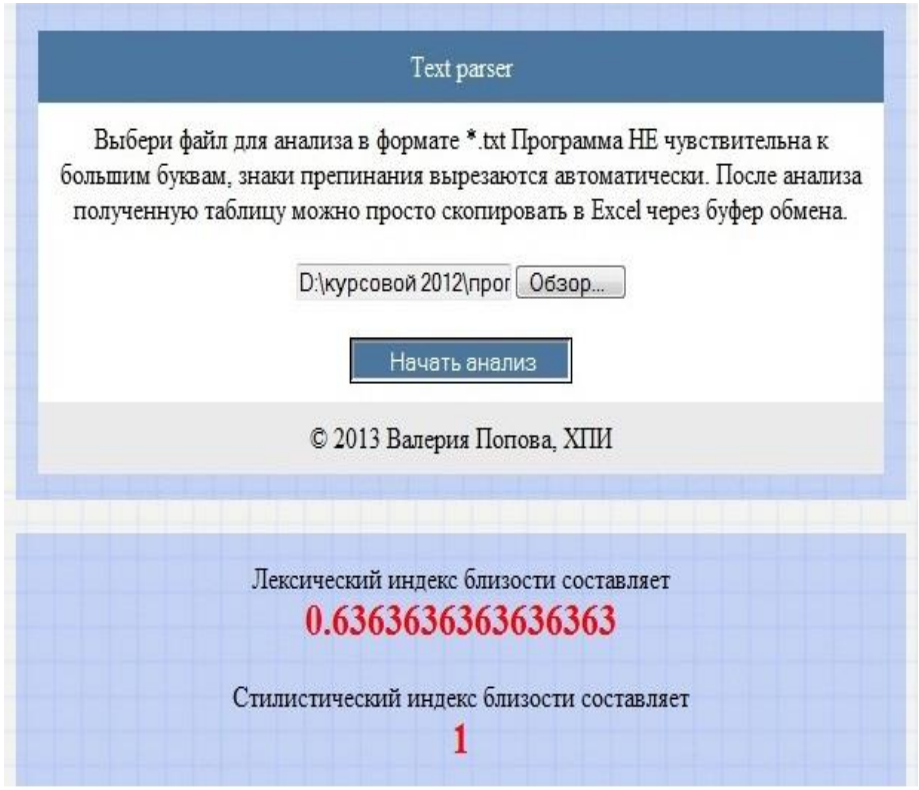
id	word	Добавить поле
1	ль вы	
2	певца любви	
3	певца своей	
4	своей печали	
5	скоро ль	
6	вздохнули ль	
7	встречали вы	
8	слыхали ль	
9	березы распустятся	
10	будет гостья	
11	в лесах	
12	в пустынной	
13	в час	
14	весне поразведать	
15	взор его	
16	взор исполненный	
17	внимая тихий	
18	встречая взор	
19	вы в	
20	вы вздохнули	
21	вы внимая	
22	вы Встречали	
23	вы за	
24	вы юношу	
25	глас ночной	
26	глас певца	
27	гостья дорогая	

Program screens of block Parser

Start of work



Result of work



Matching list of lexical items and syntactic constructions

Совпадения

Единицы

буря
зверь
как
небо
она
снежные
то

Двойки

буря мглою
вихри снежные
зверь она
как зверь
кроет вихри
крутя то
мглою небо
небо кроет
снежные крутя
то как

Тройки

буря мглою небо
вихри снежные крутя
как зверь она
кроет вихри снежные
крутя то как
мглою небо кроет
небо кроет вихри
снежные крутя то
то как зверь

Черверки

буря мглою небо кроет
вихри снежные крутя то
кроет вихри снежные крутя
крутя то как зверь
мглою небо кроет вихри
небо кроет вихри снежные
снежные крутя то как
то как зверь она

Main results and conclusions

1. Based on analysis of the subject area defined range of tasks that require linguistic expertise.
2. A review of existing systems that implement automated processing of texts in order to reduce the load of intellectual expert linguists.
3. Developed linguistic model of the text, which takes into account the statistical characteristics of the author's text.
4. For the storage and use objects of author's linguistic style structure designed lexical database language of Pushkin.
5. Developed an algorithm for determining the adjacency the texts with which the indices of lexical and stylistic adjacency texts for adoption expert opinion expert.
6. Implemented software for text recognition.

Thank You !

kanichshevaolga@gmail.com