Morphology and syntax in a problem of semantic clustering.

D. V. Mikhailov and G. M. Emelyanov
Yaroslav-the-Wise Novgorod State University

The purpose.

To develop *a mathematical model* for *revelation and classification* of syntactic relations on the set of semantic equivalent phrases concerning the problem of rise of syntactic analysis's accuracy for given Natural Language (NL).

Research tasks.

1) Development of conceptual model for semantic clustering of NL-texts on the basis of output of syntactic analysis.
2) Determination of problem area of NL-texts's semantic equivalence's establishment.
3) Development of mathematical model for revelation of laws of wordforms's linear coexistence.
4) Elaboration of advices for qualitative analysis of models of morphology and syntax for the tasks of processing Natural Language texts.

Semantic clustering on the basis of output of text's syntactic analysis: problem statement.

It is given :

A set G of analyzed texts of the given Natural Language.

It is required :

1) On the basis of output of syntactic analysis for each $T_i \in G$ to reveal:

   - a set $V(T_i)$ of *situations* described by T_i;
   - a set $M(T_i)$ of *objects* and/or *concepts* which are significant in situations from $V(T_i)$;
   - a ternary relation $I \subseteq G \times M \times V$, which puts in conformity to each $m \in M$, $M = \bigcup_i M(T_i)$ some situation $v \in V$, $V = \bigcup_i V(T_i)$, in which the considered object (or concept) appears concerning the given text T_i.

2) On the basis of the found relation I to reveal in G groups of texts, similar on occurrence of objects in the same situations.

Noun's syntactic context as the basis of forming text's attributes.

Definition 1. *As a noun's syntactic context let's consider the sequence of submitted words*

$$S_{ki} = \{v_1, \dots, v_{n(k,i)}, m_{ki}\}.$$

Here :

- v_1 *is a predicate word which designates some situation. This word is a verb or derivative noun from a verb;*
- m_{ki} *is a noun and designates some concept which is significant in* v_1*;*
- k *is an ID number of sequence among revealed from* T_i*;*
- $n(k,i)$ *is a quantity of submitted nouns in sequence* $\{v_2, \dots, v_{n(k,i)}, m_{ki}\}$.

Furthermore, for all $\{v_l, v_{l+1}\} \subset S_{ki}$ there is *a syntactic relation* R_q:

$$v_l R_q v_{l+1}, \dots, v_{n(k,i)} R_q m_{ki},$$

where q is a type of relation. It can be determined by the case of dependent word and by preposition, which connects the syntactically main and dependent word.

Transitivity of R_q allows to assert that every noun from $\{v_2, \dots, v_{n(k,i)}, m_{ki}\}$ designates some object (or concept) which is significant in v_1. Thus q defines the role of this object concerning v_1.

The Formal Concept Analysis and conceptual clustering.

Let's consider the set G of analyzed texts as a set of formal objects. Thus the set M of objects (or concepts) appearing in texts from G is a set of formal attributes. The set V of situations in which these objects (or concepts) appear is a set of formal attributes's values.

To the relation $I \subseteq G \times M \times V$ the formal context

$$K = \big(G, M, V, I\big)$$

is put in conformity.

Definition 2. *A Formal Concept (FC) is called a pair* (A, B)*:* $A \subseteq G$*,* $B \subseteq M \times V$*,* $A = B'$*,* $B = A'$*, where*

$$A' = \big\{(m, v)\colon m \in M,\ v \in V \bigm| \forall\, T_i \in A\colon m(T_i) = v\big\},$$
$$B' = \big\{T_i \in G \bigm| \forall\,(m, v) \in B\colon m(T_i) = v\big\}.$$

Definition 3. *A FC* (A_1, B_1) *is called a subconcept for the FC* (A_2, B_2)*, if* $A_1 \subseteq A_2$*, and* $B_2 \subseteq B_1 : (A_1, B_1) \leq (A_2, B_2)$*. Thus* (A_2, B_2) *is a superconcept for the FC* (A_1, B_1)*, and relation* $\leq$ *is an order relation for FCs.*

Definition 4. *A set* $\Re(G, M, V, I)$ *of all FCs for* K *together with the relation* $\leq$ *is called the Formal Concept Lattice.*

Remark. *Each word in* M *is represented together with the preposition connecting it with the syntactically main word.*

Syntactic relations concerning the situation of Natural Language usage.

Definition 5. *A situation of Natural Language (NL) usage is the description of human's social experience by means of this NL.*

Formally the language context accumulated by some such situation can be represented by a triple:

$$S = (O, R, T),$$

where O is a set of objects which participated in S, R is a set of relations between objects $o \in O$, T is a set of description forms for S in the given.

Let's assume that T is a set of NL-phrases from initial set G and each of them describes one situation of reality (relative to the language context of S).

According to arbitrariness of R, let's assume that it *consists* of *submission relations* within the frameworks of *noun's syntactic context.*

So all NL-phrases from T are strictly synonymic and

$$O = \bigcup_{T_i \in T} \big\{ M(T_i) \cup V(T_i) \big\}.$$

Here $V(T_i)$ and $M(T_i)$ are contain, accordingly, verbal designations for S and for objects (or concepts) which are associated with the situation S.

According to the definition, S is the full and independent context description.

So we have a problem:

Problem 1. *On the basis of NL-phrases from T it's necessary to form R by consideration the relations between $o \in O$ as attributes of them relative to S.*

Syntagmatic dependences as a basis of revealing syntactic relations.

Let T is a set of phrases of given Natural Language (NL) which describe some situation S. Let's consider $T_i \in T$ as a set of symbols.

For each $T_i \in T$ is true:

$$T_i = T_i^C \cup T_i^F,$$

where T_i^C is a common invariant part for all $T_i \in T$, T_i^F is an inflectional part.

Let W_{ij} is the alphabetic structure of word, j is its ID number in NL-phrase.

Then

$$W_{ij} = W_{ij}^C \cup W_{ij}^F,$$

where $W_{ij}^C \subset T_i^C$ is the invariant part and $W_{ij}^F \subset T_i^F$ is the inflectional part of word's alphabetic structure. By means of T_i^F the syntagmatic dependences are expressed.

Definition 6. *Syntagmatic dependences define linear coexistence of wordforms and are set by syntactic relations.*

So, by pairwise comparison of W_{ij} from different T_i it is required to find:

1) W_{ij}^C and W_{ij}^F for each W_{ij} when $\left|W_{ij}^C\right| \to \max$;
2) The syntactic relation R_q which defines combination admissibility for the inflections having alphabetic structures W_{ij}^F and W_{ik}^F, $k \neq j$.

Linear structure's model for Natural Language phrase.

Let T is a set of synonymic phrases and J is an index set for invariant parts of all words used in all phrases from T.

Definition 7. *Let L is the ordered set of indexes $j \in J$ of invariant parts of words presented in $T_i \in T$. We say that L is the linear structure's model for T_i.*

Let $h(j, L(T_i))$ is a position of index j in the given model $L(T_i)$.

Then the set of links relative to $L(T_i)$ can be defined as

$$D : T_i \to \Big\{ \Big(h\big(j, L(T_i)\big),\ h\big(k, L(T_i)\big) \Big) \ :\ j \neq k \Big\}.$$

Definition 8. *A link*

$$d_{qi} = \Big(h\big(j, L(T_i)\big),\ h\big(k, L(T_i)\big) \Big)$$

is acceptable for the model $L(T_i)$ if $\exists\, \{T_l, T_m\} \subset T$, $l \neq m$, and both $L(T_l)$ and $L(T_m)$ contain either $\{j, k\}$ or $\{k, j\}$ as a subsequence.

Let's assume, that for every $T_i \in T$ all $d_{qi} \in D(T_i)$ are satisfy the Definition 8.

Definition 9. *It is considered that model $L(T_i)$ is projective relative to the set of syntactic relations for T if*

$$\sum_{q=1}^{|D(T_i)|} \Delta_{qi} \leq \big|L(T_i)\big|, \text{ where}$$

$$\Delta_{qi} = \big|h(j, L(T_i)) - h(k, L(T_i))\big|.$$

Classifying syntactic relations on the basis of syntagmas graph.

Let $\bigcup_i D(T_i)$ is the set of links acceptable for all linear structures's models $L(T_i)$ of synonymic phrases T_i describing some situation.

Let's assume also that models are defined on some index set J. At acceptability of the link for $\{j,k\} \subset J$ the pair (j,k) corresponds to the single syntagma.

Definition 10. *A set of pairs (j,k), grouped by some index k common for them, is an element of the set V^J of nodes of graph (V^J, I^J) of syntagmas. Some sets E_1 and E_2 which are members of V^J will be connect by an edge from I^J if $\exists\ \{j,k,m\} \subset J$: $(j,k) \in E_1$, $(k,m) \in E_2$ and $j \neq m$.*

Let

$$G^F = \left\{ f_{ij} \colon f_{ij} = \odot \left(W^F_{ij} \right) \right\},\ I^F = \{ (f_{ij}, f_{ik}) \colon s(j,k) = \text{true},\ \{j,k\} \subset J \}.$$

Here $\odot$ is the operation of concatenation sequentially implemented with a symbols of word's inflectional part. A relation s can be recursively defined on the basis of (V^J, I^J):

1) $s(j_1, j_1) = \text{true}$;

2) $s(j_1, j_2) = \text{true}$ in one of two following cases:

— $\exists\, E_1 \in V^J$: $(j_1, j_2) \in E_1$ and $\exists\, j_3 \in J$ for which $s(j_2, j_3) = \text{true}$;

— $\exists \left(E_1, E_2 \right) \in I^J$ such that $\exists\, j_3 \in J$ and for this index we have $(j_1, j_3) \in E_1$, $(j_3, j_2) \in E_2$ and $s(j_3, j_2) = \text{true}$.

To the relation I^F we can put in conformity a formal context:

$$K^F = \left(G^F, M^F, I^F \right), \text{ where } M^F = G^F.$$

Let's name a context K^F as the *formal context of inflectional compatibility.*

Splintered Predicative Values.

Let S_1 and S_2 are the sets of *sequences of submitted words* where each sequence is a *syntactic context* of some noun.

Let's enter into consideration the following functions: *prep*, which puts in conformity for each word a preposition for link with a dependent word; *case* relating the case symbol to a noun; *norm* setting the conformity between a word and its initial form.

Definition 11. *A pair* $\{S_1, S_2\}$ *describes the Splintered Predicative Value (SPV) or conversive if for* $\forall\, S_{k1} \in S_1$ *can be find* $S_{j2} \in S_2$ *such that the following cases of mutual conformity* S_{k1} *and* S_{j2} *are possible.*

Case (1).

$$S_{k1} = \left\{v'_{11}, v_{k2}, v_{k3}, \ldots, v_{kn(k,1)}, m_{k1}\right\} \text{ and } S_{j2} = \left\{v'_{21}, v'_{k2}, v_{k3}, \ldots, v_{kn(k,1)}, m_{k1}\right\}.$$

Here $norm(v'_{11}) = norm(v'_{21})$ and $norm(v_{k2}) = norm(v'_{k2})$.
In general case $prep(v'_{11}) \neq prep(v'_{21})$ and $case(v_{k2}) \neq case(v'_{k2})$.

Case (2).

$$S_{k1} = \left\{v'_{11}, v'_{12}, v_{k2}, v_{k3}, \ldots, v_{kn(k,1)}, m_{k1}\right\} \text{ and } S_{j2} = \left\{v'_{21}, v'_{k2}, v_{k3}, \ldots, v_{kn(k,1)}, m_{k1}\right\}.$$

Here $norm(v_{k2}) = norm(v'_{k2})$ and $case(v_{k2}) \neq case(v'_{k2})$ (in general case).
It is necessary that:

- For S_{j2} $\exists S'_{k1} \in S_1$, $S'_{k1} \neq S_{k1}$: $\{S'_{k1}, S_{j2}\}$ satisfies the requirement of *Case 1.*
- For S_{k1} $\exists S'_{j2} \in S_2$, $S'_{j2} \neq S_{j2}$: $\{S_{k1}, S'_{j2}\}$ also satisfies the requirement of *Case 1.*

Thus SPV is a pair $\{v_{11}, v_{12}\}$, where $v_{11} = norm(v'_{11})$ and $v_{12} = norm(v'_{12})$.

Revealing of inflections for words within Splintered Predicative Values.

Let $W_{ij} \subset T_i$ is the alphabetic structure of word, where T_i is the set of symbols of some NL-phrase. Then W_k^P is the alphabetic structure of word, invariant part of which cannot be found in all NL-phrases of the given synonymic set.

Let's consider

$$T_i^{\odot} = \left\{ w_{ij} \colon w_{ij} = \odot(W_{ij}) \right\}.$$

Let's also assume that $\exists\, T_i^P \subset T_i$ which defines a sequence

$$P_i^{\odot} = \left\{ u_k \colon u_k = \odot\left(W_k^P\right), \bigcup_k W_k^P = T_i^P \right\}.$$

Lemma 1. *A sequence $P_i^{\odot}$ contains predicate word if*

$$\exists\, \{j, 0, k\} \subset L(T_i) \colon \ \{w_{ij}, u_1, \ldots, u_p, w_{ik}\} \subset T_i^{\odot},$$

where $\{u_1, \ldots, u_p\} = P_i^{\odot}$ and $p = \left|P_i^{\odot}\right|$.

Let T is a set of synonymic NL-phrases.

Lemma 2. *A word $u_k \in P_i^{\odot}$ is a member of Splintered Predicative Value (SPV) if $\exists\, T_j \in T \colon L(T_j) \neq L(T_i)$ and $u_k \in P_j^{\odot}$, where $P_j^{\odot}$ also satisfies Lemma 1.*

Here $\nexists\, T_k \in T$ for which $P_k^{\odot} \subset P_i^{\odot}$, $L(T_k) \neq L(T_j)$ and $L(T_k) \neq L(T_i)$.

Let $P_i^{\odot\prime}$ is a sequence of words, each of which satisfies the condition of *Lemma 2*.

Theorem 1. *For forming a formal context of inflectional compatibility at SPV's or conversive's presence it is necessary and enough to find the set $T' \subset T$:*

$$T' = \left\{ T_i \colon |P_i^{\odot\prime}| \to \max \right\}.$$

A formal context of inflectional compatibility at SPV's presence.

Let (V^J, I^J) is a graph of syntagmas and J is an index set on which the linear structures's models $L(T_i)$ for synonymic phrases from T are defined. Let's consider

$$I_1^J = \left\{(j,k)\colon \exists\, E \in V^J, (j,k) \in E\right\}.$$

By means of I_1^J an objects and attributes in *formal context of inflectional compatibility* are related. Let's name a structure $\left(V_1^J, I_1^J\right)$ as the precedent tree for T, where $V_1^J = J$.

Let $P_i^{\odot\prime}$ is a sequence of words, each of which satisfies the condition of *Lemma 2* and $T' \subset T$ satisfies the condition of *Theorem 1*.

For $\forall\, u_k \in \bigcup_i P_i^{\odot\prime}$, where $T_i \in T'$, its invariant and inflectional parts are formed by a comparison of alphabetic structure of u_k with each $u_j \in \bigcup_l P_l^{\odot}\colon T_l \in (T \setminus T')$. Here $\forall\, P_l^{\odot}$ contains a words for which an invariant part was not found initially.

Here for alphabetic structure of words it is necessary, that

$$2\left|W_k^C\right| > \left|W_k^F\right| + \left|W_j^F\right|,$$

where by means of C and F an invariant and inflectional parts of word are designated.

A tree $\left(V_1^J, I_1^J\right)$ can be transformed as follows:

- Root changes from $k = 0$ to the value of k for $u_k \in P_i^{\odot\prime}$ with a maximal occurrence in different NL-phrases from T;
- Right subtree re-hangs to the node j for $u_j \in P_i^{\odot\prime}$ with least occurrence;
- In a pair $\{u_l, u_m\} \subset P_i^{\odot\prime}$ a child node corresponds to the word with a lesser occurrence.

Experimental approbation : initial data.

Test question (in Russian):

«Каковы негативные последствия переобучения при скользящем контроле ?»

The received variants of correct answer:

«Нежелательное переобучение приводит к заниженности эмпирического риска.»
«Нежелательное переобучение, следствием которого является заниженность эмпирического риска.»
«Заниженность эмпирического риска является следствием нежелательного переобучения.»
«Заниженность эмпирического риска, являющаяся следствием нежелательного переобучения.»
«Эмпирический риск, заниженность которого является следствием нежелательного переобучения.»
«Эмпирический риск, заниженный вследствие нежелательного переобучения.»
«Эмпирический риск, к заниженности которого ведет нежелательное переобучение.»
«Риск, заниженный как следствие переобучения.»
«Эмпирический риск по причине, обусловленной нежелательным переобучением, может оказаться заниженным.»
«Эмпирический риск в силу обстоятельств, связанных с нежелательным переобучением, может оказаться заниженным.»
«Эмпирический риск по причине, вызванной нежелательным переобучением, может быть заниженным.»
«Эмпирический риск, к заниженности которого приводит нежелательное переобучение.»
«Нежелательное переобучение служит причиной заниженности эмпирического риска.»
«Заниженность эмпирического риска, причиной которой является нежелательное переобучение.»
«Заниженность эмпирического риска является результатом нежелательного переобучения.»
«Нежелательное переобучение, с которым связана заниженность эмпирического риска.»
«Эмпирический риск, с переобучением связана его заниженность.»
«Заниженность эмпирического риска связана с переобучением.»
«Заниженность эмпирического риска, являющаяся результатом нежелательного переобучения.»
«Нежелательное переобучение, результатом которого является заниженность эмпирического риска.»
«Нежелательное переобучение, результат которого есть заниженность эмпирического риска.»
«Нежелательное переобучение, приводящее к заниженности эмпирического риска.»
«Нежелательное переобучение, служащее причиной заниженности эмпирического риска.»
«Заниженность эмпирического риска относится к следствию нежелательного переобучения.»
«Заниженность эмпирического риска связана с нежелательным переобучением.»
«Нежелательное переобучение является причиной заниженности эмпирического риска.»
«Заниженность эмпирического риска, причиной которой служит нежелательное переобучение.»

Result : NL-phrases with maximal projectivity and minimal quantity of words without prototypes in alphabetic structure of invariant part.

Table 1. Correct answers $T_i \in T'$.

stem	inflectional part + preposition					
заниженн	ость	ости	ость	ости	ость	ости
эмпирическ	ого	ого	ого	ого	ого	ого
риск	а	а	а	а	а	а
нежелательн	ого	ое	ого	ое	ым	ое
переобучени	я	е	я	е	ем	е
явля	естя	—	ется	ется	—	—
следстви	ем	—	—	—	—	—
служ	—	ит	—	—	—	—
причин	—	ой	—	ой	—	—
результат	—	—	ом	—	—	—
связан	—	—	—	—	а:с	—
привод	—	—	—	—	—	ит:к

Here T' is the set of NL-phrases, each of which satisfies the condition of *Theorem 1*.

Classifying the syntactic relations for final set of NL-phrases.

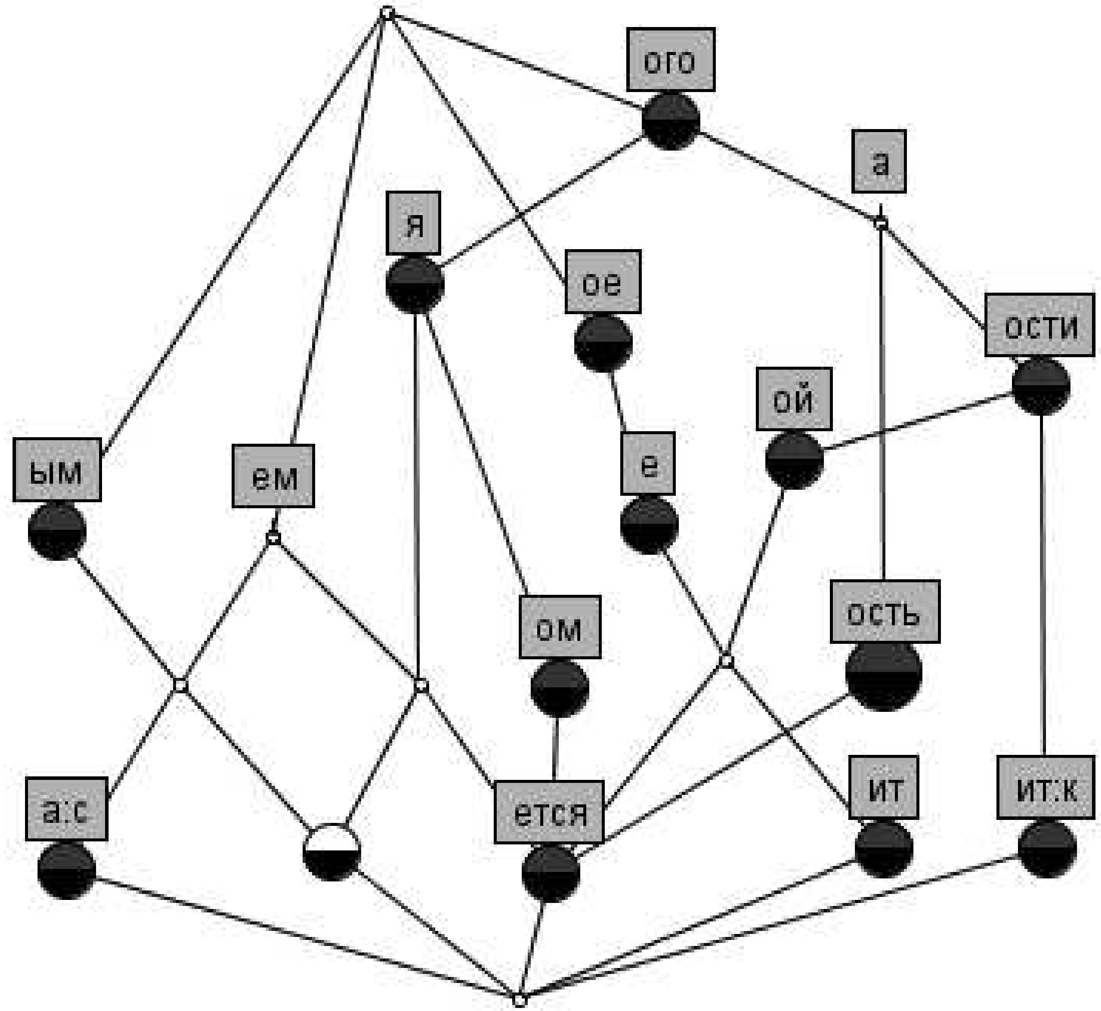
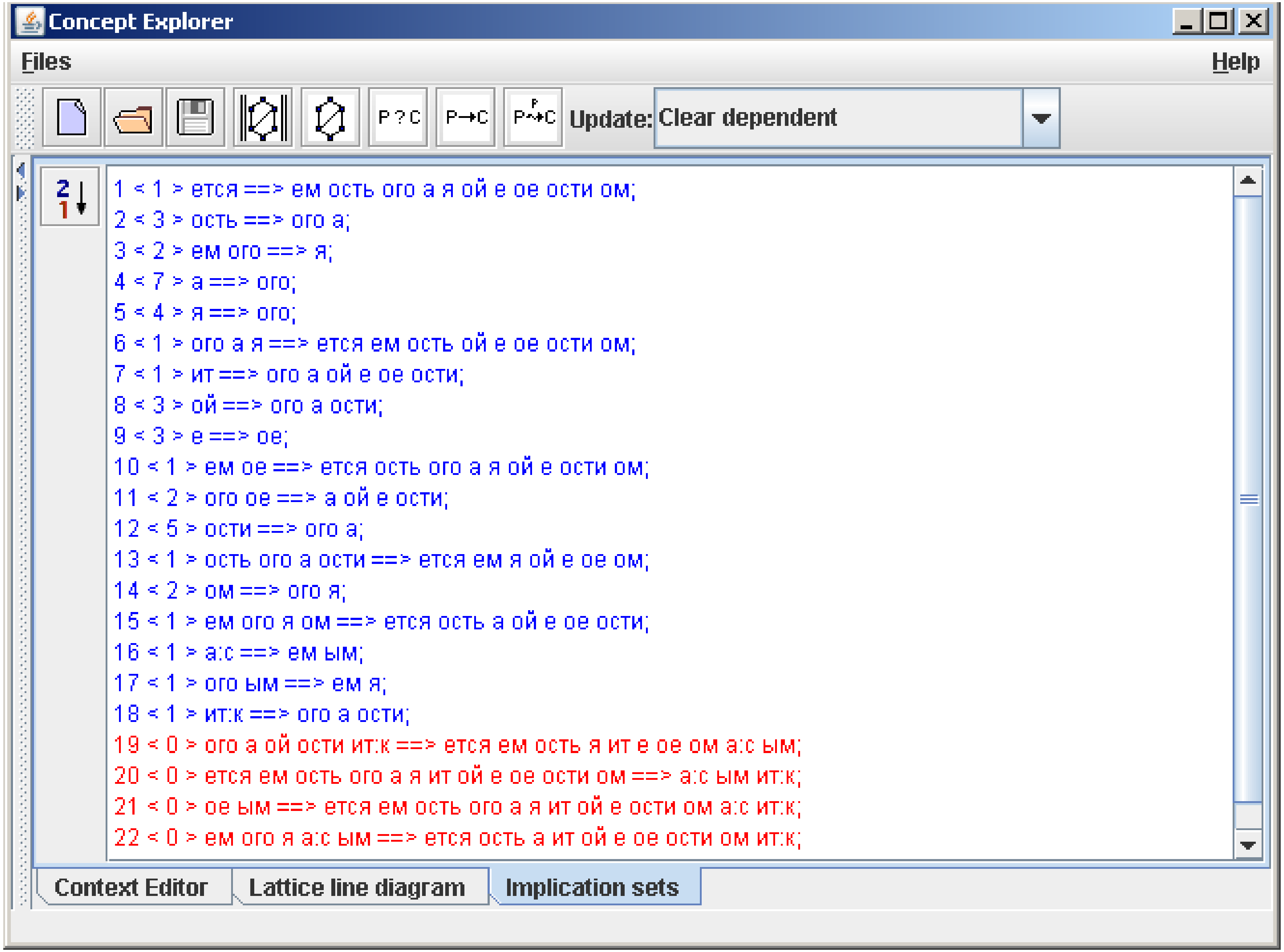


Fig. 1. Syntactic relations on the basis of inflexions's combinations.



```
1 < 1 > ется ==> ем ость ого а я ой е ое ости ом;
2 < 3 > ость ==> ого а;
3 < 2 > ем ого ==> я;
4 < 7 > а ==> ого;
5 < 4 > я ==> ого;
6 < 1 > ого а я ==> ется ем ость ой е ое ости ом;
7 < 1 > ит ==> ого а ой е ое ости;
8 < 3 > ой ==> ого а ости;
9 < 3 > е ==> ое;
10 < 1 > ем ое ==> ется ость ого а я ой е ости ом;
11 < 2 > ого ое ==> а ой е ости;
12 < 5 > ости ==> ого а;
13 < 1 > ость ого а ости ==> ется ем я ой е ое ом;
14 < 2 > ом ==> ого я;
15 < 1 > ем ого я ом ==> ется ость а ой е ое ости;
16 < 1 > а:с ==> ем ым;
17 < 1 > ого ым ==> ем я;
18 < 1 > ит:к ==> ого а ости;
19 < 0 > ого а ой ости ит:к ==> ется ем ость я ит е ое ом а:с ым;
20 < 0 > ется ем ость ого а я ит ой е ое ости ом ==> а:с ым ит:к;
21 < 0 > ое ым ==> ется ем ость ого а я ит ой е ости ом а:с ит:к;
22 < 0 > ем ого я а:с ым ==> ется ость а ит ой е ое ости ом ит:к;
```

Fig. 2. The Duquenne-Guigues set of implications.

Revealing the morphological classes of words.

Let

$$K^F = \left(G^F, M^F, I^F\right)$$

is a *formal context of inflectional compatibility* for the set G^F, where

$$M^F = G^F, I^F \subseteq G^F \times M^F,$$

and $\mathcal{L}$ is a Duquenne-Guigues set of implictions for K^F.

Let's assume also that the syntactic context for a noun m_{ki} defines by a sequence of submitted words:

$$S_{ki} = \{v_1, \ldots, v_{n(k,i)}, m_{ki}\}.$$

Rule 1. *The Formal Concept* (A^F, B^F)*:* $A^F \subseteq G^F$, $B^F \subseteq M^F$*, corresponds to the predicate word* v_1 *in* S_{ki} *if*

$$\exists\,(Pr \rightarrow Cs) \in \mathcal{L}\colon\ |Pr| = 1$$

and $Pr \cup Cs = B^F$*. Here a presence of* $(Pr_1 \rightarrow Cs_1) \in \mathcal{L}\colon Pr \subset Cs_1$ *is permissible if and only if* $Pr_1 \cup Cs_1 = B^F$*.*

Rule 2. *The Formal Concept* (A^F, B^F) *corresponds to an adjective for the noun* m_{ki}*, relative to which the syntactic context is defined, if* B^F *is the set of attributes for some element of* G^F *and* $\nexists\,(Pr \rightarrow Cs) \in \mathcal{L}\colon Pr \cup Cs = B^F$*.*

Otherwise the Formal Concept (A^F, B^F) corresponds to some noun from $\{v_2, \ldots, m_{ki}\} \subset S_{ki}$.

Conclusions.

- The basis of lattice forming for a *formal context of inflectional compatibility* are the NL-phrases with maximal projectivity, which most exactly describe the given situation, and thus more accurately express the sense. Morphological dependences revealed on the basis of affinity of inflection type for dependent words correspond to the most probable syntactic relations for the language description of the given situation.

- The proposed *model for revelation of laws of wordforms's linear coexistence* allows to reveal automatically the best way to express some thought in the given Natural Language. That allows to minimize the errors of syntactic analysis at its usage for revealing objects and attributes.

- The developed *methodology for revelation and classifying of syntactic relations on the basis of semantically equivalent NL-phrases's set* allows to automatize the development of strategies and rules of syntactic analysis. It is *especially actual* at investigation of implementation cases for given grammatical patterns in subject-oriented text corpora. An appraisal of forming knowledge here are based on similarity measures for lattices by analogy with similarity measures between the Formal Concepts.