Bull. Nov. Comp. Center, Comp. Science, 40 (2016), 37–51

Methods for constructing natural language analyzers based on link grammar and rhetorical structure theory

T.V. Batura, A.M. Bakiyeva, A.S. Yerimbetova,
M.V. Mit'kovskaya, N.A. Semenova

Abstract. The increasing volumes of Internet information and rapid development of social networks make the problem of automated text processing more and more topical. We have studied the use of link grammar for the Kazakh and Turkish languages and considered the possibility of creating dictionaries in these languages and connecting them to the LGP system. The most interesting stage of text analysis is semantic analysis. Its main goal is to represent the meaning of texts. In this paper, we are exploring the possibility of applying the rhetorical structure theory to the Kazakh language. Some of the formal features of rhetorical relations have been described. Statements about the properties of these features have been formulated. In the future, we are planning to build a system for identifying the topic of a text and an automatic summarization system based on the results received. We believe that even partial implementation of a semantic annotation can increase the overall performance of these systems.

Keywords: link grammar, agglutinative language, semantic analyzer, rhetorical structure theory, discourse markers, rhetorical relations.

Introduction

The necessity to explore link grammar arose in connection with the problem of natural language text processing, in particular, when automatic summarization and topic identification systems were being built.

Today, many morphological and syntactic analyzers are being developed. In particular, some approaches applied to agglutinative languages are described in [1-–5]. Mostly, however, the authors consider only the morphological structure of the Kazakh and Turkish languages and conduct a comparative analysis of these languages.

The paper [6] describes a software tool for resolving the morphological ambiguities in the Tatar language. The authors chose a method based on contextual rules. Indeed, it seems the most effective because agglutinative languages have regular grammar. However, in our opinion, we can only distinguish syntactic and semantic relations at the stage of morphological analysis (as shown below), which is due to the specific features of word

formation in the languages of this type. Unfortunately, the syntax and semantic structure of the Turkic languages have not been adequately studied, which complicates the automation of these processes.

The main task of our work is to explore how the rhetorical structure theory and link grammar can be applied to construct text analyzers in the Kazakh and Turkish languages. We have chosen these languages because of an expansion of Islamic culture and the a wide occurrence of texts in these languages on the Internet.

In recent years, it has often been declared that linguistic phenomena can not be clearly understood and described out of the context, without regard to their discursive aspects [7]. The discourse is often identified with a text which consists of sentences (communicative language units) and their combinations in larger unities that are in a permanent semantic connection. In other words, the discourse is not only a coherent sequence of sentences opposed to an isolated sentence, but also a certain semantic unity that has semantic connectedness [8] and as a result contains knowledge about the world and the situation, as well as social and other types of knowledge.

There have been some attempts to use discourse analysis for solving various problems of computational linguistics. A detailed review of the literature presented in [9] shows that in most cases the discourse analysis is able to improve the quality of automatic systems by 4-44%, depending on the specific problem. Research in this field for the English language has reached a sufficiently high level. There is not enough research for Russian [9–11], and for the Kazakh language such studies have not been conducted yet.

1. Link Grammar Parser

The Link Grammar Parser is a syntactic parser based on link grammar. It was created by Daniel Slitor and Davy Temperley. A detailed description of Link Grammar Parser can be found in [12]. For a given sentence, the system assigns a syntactic structure which consists of a set of labelled links connecting pairs of words. The main idea of link grammar allows working with the original theory of syntax and morphology at the same time.

Such an approach considers words as blocks with outgoing connectors. They are of different types, and can point to the right or to the left. A left-pointing connector connects with a right-pointing connector of the same type on another word. Two connectors together form a "link". The right-pointing connectors are marked by "+", and the left-pointing connectors are marked by "−".

Global rules. Words have rules about how their connectors can be linked/joined, that is, rules about what constitutes a valid use of a word. A valid sentence is the one in which all the words are used in the right way,

valid according to their rules and also satisfying certain global rules. In other words, in addition to the rules listed in the dictionary, there are two global rules that govern word connection: the planarity rule and connectivity rule. Let us explain what they are.

The planarity rule requires that links should not cross. For example, the way of connecting the four words ("tree" to "book" and "car" to "pizza") shown in Figure 1 would be illegal. The parser will just not find such links.

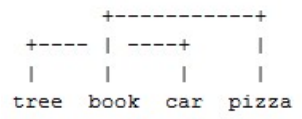

Figure 1. Planarity rule

The connectivity rule imposes the following restriction: all the words in a sentence should be connected directly. The way of connection between these four words shown in Figure 2 would be illegal.

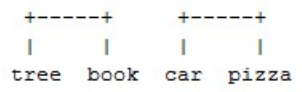

Figure 2. Connectivity rule

Parsing algorithm. Parsing is implemented in analogy to assembling a jigsaw puzzle (symbolizing the parsed sentence) from puzzle pieces (representing individual words). A language is represented by a dictionary having words and a set of allowed "jigsaw puzzle shapes" that the words can have. This "shape" is shown by a "connector," which we have mentioned previously. Thus, a common noun may have the connectors D- & S+ indicating that it may connect to a determiner which on the left ("D-") and the subject on the right ("S+"). The determiner indicates whether the noun refers to a definite or indefinite element of a class, a closer or more distant element, an element belonging to a specified person or thing, a particular number or quantity, etc. Besides, parsing indicates that the S+ connector can be attached to the S- connector, forming an "S" link between the two words.

A given word may have dozens or even hundreds of allowed "puzzle-shapes" (determined as "disjuncts" here). For example, many verbs can be optionally transitive, which makes the O+ connector optional; such verbs might also take adverbial modifiers (E connectors) which are intrinsically optional. Therefore, a part of parsing also involves selection of a single unique disjunct for a word; the final parse must connect all connectors for that disjunct.

Dictionary entries. A dictionary entry includes a word followed by a colon followed by a connector expression followed by a semi-colon. The dictionary consists of a series of such entries. Any number of words, separated by spaces, can be inserted in a list; they will then conform to the linking requirement. A connector name should include one or more capital letters (any number may be used) followed by any number of lower-case letters mixed with the wild-card character "$*$" and terminated by "$+$" , "$-$" or "$".

At the moment, there are plug-in dictionaries for the English, Russian, Persian, Arabic, German, Lithuanian, Vietnamese, and Indonesian languages. We are developing dictionaries for the Kazakh and Turkish languages.

2. Links indicating morphological features of words

Links indicating morphological features of words contain information about word formation and word combination. As the Turkish and Kazakh languages are agglutinative, the formation of new words and word forms is performed by the successive addition of affixes.

There are various types of affixes for different parts of speech [13–16]. Each type corresponds to a specific morphological feature (a singular or plural form of a noun, a person or tense of a verb, etc.) and can be associated with a connector linked to the previous suffix or stem. Then the sequential addition of morphological links allows to simulate the process of word formation. The connector may point out from the last affix to the previous, and then to the stem. For example, the verb *"to read"* is formed in the Turkish language as follows:
okuyorlar = oku + yor + lar, where
oku is a stem;
yor is a tense suffix, indicating that the action takes place at the moment;
lar is a plural suffix.

Plural nouns in the Turkish language are characterized by the presence of *–lar/-ler* affixes, attached directly to the stem of the word. These affixes can be described as <lar, ler>: {Np-}. Similar suffixes are present in the Kazakh language: <лар, лер, дар, дер, тар, тер>: {Np-}. Therefore, the connector "Np+" is necessary for the stems in the dictionary.

The possessive form of nouns and pronouns in the Turkish language is characterized by the presence of affixes *-m, -ım, -im, -um, -üm; -n, -ın, -in, -un, -ün; -sı, -si, -su, -sü, -ı, -i, -u, -ü, -mız, -miz, -muz, -müz, -ımız, -imiz, -umuz, -ümüz, -nız, -niz, -nuz, -nüz, -ınız, -iniz, -unuz, -ünüz, -ları, -leri.* A similar situation is observed in the Kazakh language. Such affixes (depending on the person) are described by means of the links Np1-, Np2-, Np3-, Pp1-, Pp2-, Pp3-.

Similarly, we can describe noun case suffixes: “Nn” for nominative; “Ng” for the genitive; “Nd” for the dative; “Na” for the accusative; “Ni” for the instrumental; “Nl” for the locative; and “Nb” for the ablative.

For example, the noun “book” (*someone's*) in the Turkish language is formed as follows: *kitabını = kitab + ı + nı*, where *kitab* is the stem; *ı* is a possessive suffix; *nı* is an accusative suffix. Then, according to the notation, we obtain the following set of morphological connectors in the dictionary: < kitab > : {Np3 +}; < I > : {Np3-} & {Na +}; < Nı > : {Na-}.

In the Kazakh and Turkish languages, affixes are usually connected to each other in a particular sequence. First, we have a word stem, then a plural suffix, then a possessive suffix, then a person suffix, and finally a case suffix. By this rule, we write down connectors in the dictionary.

For example, the noun *friends (our)* in the Kazakh language: *достарымызға = дос + тар + ымыз + ға*, where *дос* is a stem word; *тар* is a plural suffix; *ымыз* is a person suffix; *ға* is a dative suffix. We obtain next representation in the dictionary <дос>: {Np+ }; <тар>: {Np-} & {Pp1+}; <ымыз>: {Pp1-} & {Nd+}; <ға>: {Nd-}.

Some formative affixes allow to form adjectives from the nouns <лы, лі, ды, ді, ты, ті, сыз, сіз, дай, дей, тай, тей, лық, лік, дық, дік, тық, тік, ғы, гі, қы, кі >: {As-}, for example, *ай (month) – айлық (monthly)*. Other formative affixes create verbs from nouns and adjectives <да, де, та, те, ла, ле, а, е, ар, ер, қар, кер, ғар, гер >: {Vna-}. Verbal suffixes, moreover, require an adherence to the suffix *–y*, forming the verb infinitive, or a personal affix. For example, the verb *бастау (start)* is formed from the noun *бас (a start)* and has the following description in the dictionary: <бас>: {Vna+ }; <That>: {Vna- } & {V+ }; <Y>: {V-}.

Nouns derived from verbs are characterized by the presence of the affixes <шы, ші, ғыш, гіш, қыш, кіш, ма, ме, ба, бе, па, пе>: {Sv-}, for example, *оқу (learn) – оқушы (learner)*. Nouns formed from nouns are characterized by the affixes <кер, гер, лас, лес, дас, дес, тас, тес, лық, лік, тық, тік, дық, дік, шы, ші >: {Ss-}, e. g. *ғарыш (space) — ғарышкер (spaceman)*.

There are seven most commonly used verb tenses in the Kazakh language [15], each with its typical suffixes (see Table 1).

Figure 3 shows the parsing of the sentence containing the verb with the goal future tense suffix: *Ол үйде кешке болады. (He will be at home in the evening.)*

3. Links indicating syntactic features of words

We denote the syntactic functions of words in a sentence by capital Latin letters. We have identified the following basic connections in the Kazakh and Turkish languages: AS is an attribute of a subject; AO is an attribute

Table 1. Verb tenses in the Kazakh language

Verb tenses	Suffixes and connectors
Aorist Past	<ыпты, іпті>: Vas+
Past Perfect	қан, ған, кен, ген>: Var+
Past Simple (categorical)	<ты, ті, ды, ді>: Vac+
Present Simple	<п, ып, ір, а, е>: Vr+
Future Transitive	<ады, еді>: Vft+
Future Perfect Indefinite	<ар, ер>: Vfs+
Goal-oriented Future	<мақ, мек, пақ, пек >: Vfg+

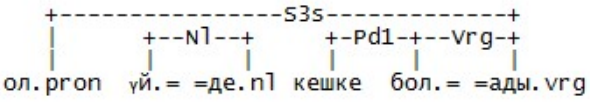

Figure 3. Parsing of the sentence containing the verb with the Goal-oriented Future tense suffix

of an object; E is an adverbial modifier; J connects a postposition and a noun; OV is a direct object; OJV is an indirect object; and S connects a subject and a predicate.

If we consider syntactic features of words in a sentence, we can associate each part of speech with a formula of possible connectors: a noun may act as a subject connected to an attribute, a verb has to be at the end of a sentence, etc. Here is an example of a sentence structure in the Turkish language: $<$N_S$>$: {AS-} & {OV+} & S+.

Besides, a noun may act as an object with an attribute on the left and predicate on the right. Such structure is generally described by the formula $<$N_O$>$: {AO-} & {OV+} & {OJV+}.

Another example shows that a verb can act as a predicate sentence, which on the left may be the subject, (direct or indirect) object or adverbial modifier: $<$V_P$>$: {EI-} & {OV-} & {OJV-} & {S-}.

At the same time, the connector AI+ should be in the description of an adjective as the necessary pair of AI-, and the connector EI+ should be in the description of an adverb as the necessary pair of EI-. Otherwise, the link will not be found.

Let us consider the sentence *Адамдар алма жеді. (People ate an apple.)*. The parser identifies two syntactic (S3p, OV) and two morphological (Np, Va3p) links. An example of this parsing is shown in Figure 4.

Another example is the sentence with an indirect object: *Иттер мысықтардың артынан қуды. (Dogs chased the cats.)*. Figure 5 shows that the parser identifies three syntactic (S3s, OJV, J) and four morphological (Np, Va3s) links.

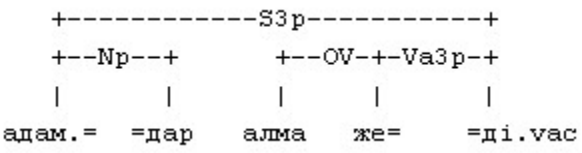

Figure 4. The sentence with the direct object

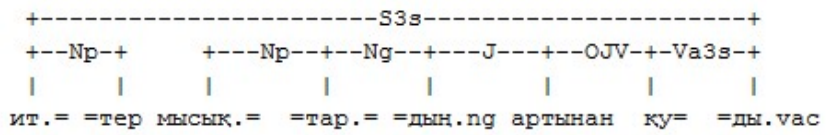

Figure 5. The sentence with the indirect object

It should be noted that the syntactic links can occur not only between words, but also between groups of words, for example, in compound verbal predicates, compound nominal predicates, participles, etc.

4. Model of the semantic parsing of sentences

4.1. Using the link grammar

To make a transition from morphological and syntactic to semantic links, it is more convenient to switch to the terms of predicates. We have double predicates because we use link grammar.

Thus, the syntactic links discussed in the previous section can sometimes be saved in the form of predicates: AS (adjective, noun); AO (adjective, noun); E (adverb, verb); OJV (Nd (noun) | Na (noun) | Ni (noun) | Nl (noun) | Nb (noun), verb); S (Nn | Pn), verb), etc.

Note that under this assumption the predicates OV *(x, y)* and OJV *(x, y)* contain information about verbal coordination, that is, they depend on the use of a specific case before a certain verb. In the future, we plan to study further the verbal coordination in the Kazakh and Turkish languages. Now, we can consider the semantic predicate of possession:
OF (Possessor, ObjectOfPossession) = OF (Ng (noun) | Pg (pronoun), Np3 (noun)).
The predicate OF *(x, y)* describes, for example, the phrase:
kadının elbisesi ("women's dress", i.e. dress which belongs to the woman), where *kadın* is a stem of a word ("female"); *ın* is a genitive suffix; *elbise* is a stem of a word ("dress"); *si* is a possessive suffix.

Consider the sentence *Ben kardeşin kitabını okuyorum. (I am reading the brother's book.)*. Let us write this sentence with the help of the predicates: READ (*ben*, OF *(kardeşin, kitabını)*).

The predicate OF emphasizes possessive pronouns. Figure 6 shows a parsing example, *Менің қарным ашқан жоқ. (I am not hungry.)* containing the first person possessive pronoun (the link OF1 is responsible) and the negative form of the verb (the link VN is responsible).

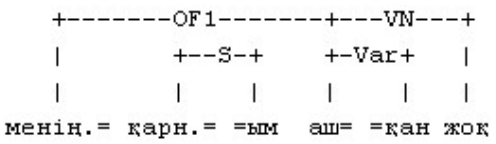

Figure 6. Possessive pronouns in the Kazakh language

An example sentence parsing with a possessive pronoun is shown below *(Senin ne istedigini bilmiyorum. — I don't know what you want.)*

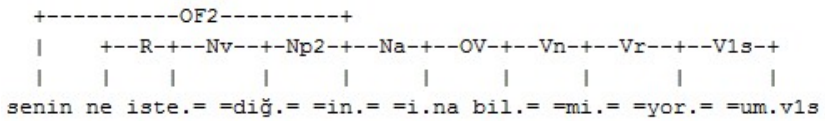

Figure 7. Possessive pronouns in the Turkish language

Semantic predicates of place LOC (verb, adverb) and time of action TIME (verb, adverb) are interesting from the perspective of further research.

The predicate FOR (Ng (noun) | Pg (pronoun), postposition) describes a combination of a postposition *"icin"* with a noun or pronoun in the genitive case.

4.2. Using the rhetorical structure theory

Rhetorical Structure Theory (RST) is one of the best known theories of text organization [17]. According to it, a text is initially divided into non-overlapping fragments called elementary discourse units (EDU). For example: *Mark Zuckerberg is a programmer.* ***However,*** *he retrained as a businessman.* It can be divided into two parts.

$$[\textit{Mark Zuckerberg is a programmer}]^N,$$
$$[\textbf{\textit{However,}}\ \textit{he retrained as a businessman}]^S.$$

Then, elementary discourse units are joined between themselves by rhetorical relations. These parts are the elements that comprise the larger fragments of texts and whole texts. Each fragment has a particular role vis-a-vis other fragments. Text connection is formed by relations modeled between fragments within the text. The set of rhetorical relations is set in advance. In the study [18], it consists of 21 relations; in other studies [10, 19], it has 27

and even 47 elements, respectively. The ultimate aim of a discourse parser is to build a tree structure of a text and show how the parts of the text are related to each other.

Two types of elementary discourse units are defined in RST. The nucleus is considered the most important part of the text, while satellites clarify nucleus and are secondary. The nucleus contains basic information, and satellites contain additional information about the nucleus. A satellite is often incomprehensible without the nucleus. On the contrary, sentences where satellites have been removed are understandable to some extent. If a nucleus is omitted or changed, the meaning of the text and relation varies significantly [20].

There are symmetrical (multi-nuclear) and asymmetrical (mononuclear) relations.

Symmetrical relations can connect any number of discourse units without differentiating between them. For example, *List* is a multi-nuclear relation where the elements are listed but not compared.

In asymmetrical relations, two discourse units have different status, i.e. the relation between them is directed. An arrow points at the nucleus of the relation, and the beginning of the arrow points at the satellite. Thus, the Condition relation can be described as follows. The satellite represents a hypothetical event, the future or unrealized situation. For instance, let us considerer the sentence **Егер** *көп еңбектенсең, бір нәрсе шығады.* (***If*** *you suffer for a long time, you will succeed.*) It can be presented by the following components: $[$*Егер көп еңбектенсең*$]^S$, $[$*бір нәрсе шығады*$]^N$ ($[$*If you suffer for a long time*$]^S$, $[$*you will succeed*$]^N$). Then semantic parsing will look as follows (see Fig. 8).

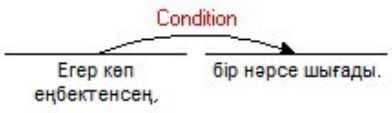

Figure 8. Semantic parsing in terms of the rhetorical structure theory

All rhetorical relations may be presented by pairs of three types:

1. nucleus-nucleus $< \mathrm{N}, \mathrm{N} >$ (Nucleus-Nucleus) -– a symmetrical rhetorical relation;
2. nucleus-sattelite $< \mathrm{N}, \mathrm{S} >$ (Nucleus-Satellite) -– an asymmetrical rhetorical relation where the nucleus precedes the satellite;
3. sattelite-nucleus $< \mathrm{S}, \mathrm{N} >$ (Satellite-Nucleus) -– an asymmetrical rhetorical relation where the satellite precedes the nucleus;

The main problem is that it is quite difficult to determine the definite rhetorical relation which connects the discourse units. The features of each relation must be described formally, and on this basis the correspondence may be established. Then we could use the rhetorical structure theory in the systems of automatic text analysis.

As it is noted in [10], rhetorical relations can be considered as predicates with properties indicating certain differentiating features. For some rhetorical relations markers can be detected. For example, *Ivan arranged a party to have fun.* The rhetorical relation corresponds to the relation *Purpose*, where the satellite with the marker "to" is attached to the nucleus.

Research of discourse markers is one of the most popular fields of discourse analysis [21, 22]. The conjunctions (*when, because, and so on*) are often related to a class of discourse markers. Besides, this class may include the markers of the speaker's mental processes (*here, well, so to speak*), markers of control over the addressee's mental processes (*you know, you see*), and others [23]. It should be noted that the discourse markers in the Kazakh language have not been adequately studied, although the principles of discourse analysis do not depend on a language and can be used for Russian and other languages.

Below you can see the markers, corresponding rhetorical relations, and some examples (see Table 2).

It is impossible, however, to characterize most of rhetorical relations by the presence of certain discourse markers. Moreover, the markers themselves are nonuniversal features since they are expressed differently in different natural languages. Therefore, other features must be selected for a clearer description of rhetorical relations. For instance, we can take as features the classifying parameters described by M. Luvers [24]. He identified the parameters most commonly used to describe the relations of cohesion and coherence. Cohesion is the structural connectedness of a text, and coherence is the meaningful connectedness of a text. Mostly, "coherence" refers to the content of the text; it is the organisation of the content of the text as a whole, when the communicative situation itself and the set of knowledge of the sender and recipient are important. "Cohesion" applies to the structural organization of the text and it is responsible for the connection of the text units by means of particular language levels. In other words, coherence is an attribute of the text and cohesion is an attribute of the text elements.

M. Luvers identified four types of parameters: the type of relations, polarity, direction and reflection of relations in the real world.

The first of these parameters is the type of relations. Relations can be of three types: TYPE = C, T, A, where C is causal, T is temporal, and A is additive. Causality contains an indication of time and reason; temporality includes only time; and additivity does not contain any indications.

The polarity of relations means that they can be positive or negative:

Table 2. Examples of markers and sentences with them

Marker	**Rhetorical relation**	**Example of sentence**
moreover (оған қоса)	Elaboration	The house looked well. *Moreover*, the price was suitable. Үй жаман емес көрінді. *Оған қоса*, бағасы да тиімді болды.
because (себебі)	Evidence	She is very clever, *because* she studied hard, scrutinizing every task. Ол өте білімді, *себебі* бар ынтамен оқып, барлық тапсырмаға ұғынып жатты.
if..., then (егер. . . , онда)	Condition	*If* you are not going to study, *then* it will be difficult. *Егер* оқымасаң, *онда* қиын болады.
although (дегенмен)	Concession	It turned out to be interesting, *although* some scientists say laughing that it was just a coincidence. Қызық шығыпты, *дегенмен* кейбір ғалымдар келемештеп, оны жай кездесоқ оқиға деп айтып жатыр.
for (үшін)	Purpose	I bought some coloured pensils *for* drawing a picture. Мен сурет салу *үшін* бірнеше түсті қарындаштар сатып алдым.

POL = P, N. Positivity means that the first situation develops into another situation attached to it. On the contrary, negativity assumes that the expected connection of the situations terminates, and an opposition emerges.

Direction can be forward, backward and bi-directional depending on the order of events mentioned in the text: DIR = B, F, BD.

The reflection of relations in the real world can be considered on two levels: between events and between speech acts. M. Luvers refers the former level to semantics and the latter, to pragmatics. Semantic relations are divided into object-matter and subject-matter: SEM = O, S. Pragmatic

relations are divided into intentional and presentational: PRAG = IN, PR.

It should be noted that the last-mentioned, forth parameter (reflection of relations in the real world) is the least studied in the literature and the most difficult to formalize. Therefore, it was decided to consider only the first three and to leave the study of the forth parameter for the future.

In addition to these three parameters, some rhetorical relations can be described using the LGP links introduced in the previous section. Examples of the description are presented in Table 3.

Table 3. Description of the relations using the LGP links and other features

Rhetorical relation	Possible marker	Description using the LGP links	Type	Polarity	Direction
Elaboration *Бөлшектеу*	moreover *(оған қоса)*	E+ or (Xl- & Xr+ & (E+ or E-)) or (Xr+ & Xl- & OJV+) or (Xr+ & Xl- & E-)	add	pos	F
Evidence *Дәлел*	because *(себебі)*	(J+ or E+) & (E- or J- or (Xr+ & Xl- & OJV+) or (Xl- & Xr+ & E+))	caus	pos	B/F
Condition *Шарт*	if. . . , then *(егер . . . , онда)*	(OV- & Xc+ & Xd- or (OV- or E+) & ((Xl- & Xr+ & E-) or (Xr+ & Xl- & OJV+)))	caus	pos	F
Concession *Көну*	although *(дегенмен)*	(OV- & ((Xr+ & Xl- & OJV+) or (Xl- & Xr+ & E-)))	caus	neg	B/F
Purpose *Мақсат*	for *(үшін)*	((J+ or E+) & (OV- or MVI-)) or (J+ & (Xr- or Xl+))	caus	pos	F

Let us formulate the following statements about the properties of these features.

Statement 1. Additivity in the backward direction is impossible.
Let TYPE = C, T, A and DIR = B, F, BD.
If R $\in$ A, then R $\in$ F or R $\in$ BD, but R $\notin$ B.

Statement 2. Bi-directional causality does not exist.
Let TYPE = C, T, A and DIR = B, F, BD.
If R $\in$ C, then R $\in$ B or R $\in$ F, but R $\notin$ BD.

Statement 3. Negative bi-directional temporality does not exist.
Let TYPE = C, T, A, DIR = B, F, BD and POL = P, N.
If R $\in$ T and R $\in$ BD, then R $\in$ P and R $\notin$ N.
The proof of these statements follows directly from the definitions of TYPE, DIR and POL.

Conclusion

Studies of natural languages involving mathematical models and methods are still of current interest in view of the rapid increase in the volume of text information. This paper describes automatic methods for the morphological, syntactic and semantic analysis of texts. For morphological and syntactic analyzers, a dictionary has been composed which includes about 2000 verbal affixes and their combinations and about 3500 affixes and their combinations for nouns and adjectives. It has been observed experimentally that such a volume is sufficient to analyze the texts on any subject. For texts in the Kazakh language, the accuracy of morphological analysis was 0.947, and the unlabeled attachment score (UAS) of parsing of simple sentences was 0.59.

The most interesting and challenging stage is semantic analysis. Its main task is to obtain a formal representation of the meaning of a text. In this paper, we have studied the possibility of applying the rhetorical structure theory and link grammar to the Kazakh and Turkish languages. The main problem is that it is quite difficult to determine the specific rhetorical relation that connects discourse units. We have made an attempt to describe formally the features of some relations. Statements about the properties of these features have been formulated.

The proposed method of the formalized description of text structure using rhetorical relations takes into account the hierarchical nature of the text, determines the conditions for combining its fragments, and therefore can be used in automatic text summarization systems. In this work, we have given only a few rhetorical relations as an example as we were not faced with the task of covering as many cases as possible. To build a full-fledged analyzer, it is necessary to take into account at least 21 relations. Undoubtedly, an increase in the number of considered rhetorical relations improves the quality of summaries.

In the future, we plan to continue the development of the automatic text processing system with the help of the proposed methods in order to test their effectiveness in solving the problem of text summarization.

References

[1] Eryigit G., Nivre J., Oflazer K. Dependency parsing of Turkish // Computational Linguistics. – 2008. – Vol. 34, No. 3. – P. 357–389.

[2] Oflazer K. Two-level description of Turkish morphology // Literary and Linguistic Computing. – 1994. – Vol. 9, No. 2. – P. 137–148.

[3] Tukeyev U. Automaton models of the morphology analysis and the completeness of the endings of the Kazakh language // Proc. Intern. Conf. “Turkic languages processing” / TurkLang-2015, Kazan, Russia, September 17–19, 2015. – P. 91–100.

[4] Zhumanov Zh.M. Writing a grammar for the syntactic analysis of the Kazakh language // Vestnik of KazNU. Series: Mathematics, Mechanics, Computer Sciences. – 2012. – No. 2 (73). – P. 71–80 (In Russian).

[5] Tukeyev U.A. Zhumanov Zh.M., Rakhimova D.R. Modeling of the semantic situations of tenses of the Kazakh language using the machine translation // Vestnik of KazNU. Series: Mathematics, Mechanics, Computer Sciences. – 2012. – No. 4 (75). – P. 99–107 (In Russian).

[6] Suleymanov D.Sh., Gilmullin R.A., Gataullin R.R. Software tool for morphological disambiguation in the Tatar language // Open Semantic Technologies for Intelligent Systems (OSTIS–2014). – Minsk, 2014. – P. 503–508 (In Russian).

[7] Prokoshenkova L.P., Geckina I.B. Discourse analysis and its role in modern linguistics // Vestnik of ChSU. – 2006. – No. 4. – URL: http://cyberleninka.ru/article/n/diskursivnyy-analiz-i-ego-rol-v-sovremennoy-lingvistike (In Russian).

[8] Temnova E.V. Modern approaches to the study of discourse // Language, Consciousness, Communication: A Collection of Articles. – 2004. – Iss. 26. – P. 24–32 (In Russian).

[9] Ananeva M.I. Kobozeva M.V. Development of corpus in Russian with the markup on the basis of rhetorical structure theory // Proc. Intern. Conf. on Computer Linguistics and Information Technologies / Dialog 2016, Moscow, Russia, June 1–4, 2016 – URL: www.dialog-21.ru/media/3460/ananyeva.pdf (In Russian).

[10] Susov A.A. Modeling discourse in terms of the rhetorical structure theory // Vestnik of VSU. Series: Philology. Journalism. – 2006. – No. 2. – P. 133–138 (In Russian).

[11] Kibrik A.A. Analysis of discourse in cognitive perspective. – Thes... doct. phylology (language theory). – Moscow, 2003. – URL: http://iling-ran.ru/kibrik/DA_cognitive_perspective@Diss_2003.pdf (In Russian)

[12] Temperley D. An Introduction to the Link Grammar Parser. – 2014. – URL: http://www.abisource.com/projects/link-grammar/dict/introduction.html#1

[13] Kessikbayeva G., Cicekli I. Rule based morphological analyzer of Kazakh language // Proc. of the 2014 Joint Meeting of SIGMORPHON and SIGFSM / Conf. MORPHFSM 2014, Baltimore, USA, June 27, 2014. – P. 46—54.

[14] Özlem İstek. A Link Grammar for Turkish. – Thes... M.S. in Computer Engineering. -– Ankara, Turkey, 2006. – URL: http://www.cs.bilkent.edu.tr/ ilyas/PDF/THESES/ozlem_istek_thesis.pdf.

[15] Kulikovskaya L.K., Musaeva E.N. Grammar of the Kazakh Language in Tables and Schemes in Comparison to Grammar of Russian. -– Almaty, 2006 (In Russian).

[16] Tukeyev U.A., Rakhimova D.R. The semantic links in automatic text processing of the Kazakh language // Vestnik of KazNTU. Series: Mathematics, Mechanics, Computer Sciences. – 2012. – No. 2. – P. 320–325 (In Russian).

[17] Mann W., Thompson C. Rhetorical structure theory: Toward a functional theory of text organization // Text-Interdisciplinary J. for the Study of Discourse. – 1988. – Vol. 8, No. 3. – P. 243–281.

[18] Ananeva M.I. Kobozeva M.V. Discourse analysis in natural language processing // Proc. 4th Russian Scientific Conf. of Young Scientists with International Participation "Informatics, Management and Systems Analysis". – 2016. – Vol. 1. – P. 138–148 (In Russian).

[19] Litvinenko A.O. Description of the discourse structure within the rhetorical structure theory: application to Russian material // Proc. Intern. Workshop "Dialog-2001" on Computational Linguistics and Its Application. – 2001. – Vol. 1. – P. 159–168 (In Russian).

[20] Kovalchuk N.V., Volodina M.S. The rhetorical structure theory as a pragmatic concept of text analysis // Bull. of the Northern (Arctic) Federal University. Series "Humanitarian and Social Sciences". – 2016 . – No. 3. – P. 107–113 (In Russian).

[21] Baranov A.G. Functional-Pragmatic Concept of a Text. – Thes... doct. phylology: 10.02.19. – Krasnodar, 1993 (In Russian).

[22] Fraser B. What are discourse markers?// J. of Pragmatics. – 1999. – Vol. 31, No. 7. – P. 931–952.

[23] Palatovskaya E.V. Discourse analysis and the rhetorical structure theory // Scientific Bull. of UNESCO Chairs KNLU. Series: Philology, Pedagogy, Psychology. – 2014. – Iss. 29. – P. 89–95 (In Russian).

[24] Louwerse M. An Analytic and Cognitive Parameterization of Coherence Relations // Cognitive Linguistics. – Cambridge, 2001. – Vol. 12, No. 3. – P. 291–316.