Functional and Cognitive Aspects in Linguistic Modelling

E. Kozerenko
Institute of Informatics Problems of the Russian Academy of Sciences, Moscow, Russia

Abstract - *The paper focuses on the issues of establishing transferable language phrase structures. The approach employed is based on generalized cognitive entities manifested in the categorial systems of the English and Russian languages and functional roles of language units in a sentence. The formalism developed for presentation of syntactic structures for the English-Russian machine translation is a variant of unification grammar and comprises about three hundred rules. A number of declarative modules of linguistic processors were designed and implemented within the framework of machine translation system"Cognitive Translator" and knowledge extraction systems.*

Keywords: machine translation, syntax, semantics, set-phrase dictionaries, machine learning, translation memory

1 Introduction

To face the problems of language structures transferability for machine translation (MT), it is necessary to consider human translation experience. Translation is a creative and sophisticated human activity, hence, producing automatically a high-quality translation of an arbitrary text from one language to another is a task too far from its complete implementation. However, for simpler tasks, such as acquiring information on the Web, getting acquainted with subject domain information, etc., rough translation output without post editing can be quite adequate. One of the domains where MT works best is scientific discourse. Perhaps, it can be accounted for the regularity of syntactic structures which is required by the functional style of scientific prose.

Of the three forms of translation performed by man: written translation, consecutive interpretation and simultaneous interpretation, the one which is nearest to the real-time machine translation is simultaneous interpretation (SI). Therefore, the recommendations for SI are of prime interest to MT designers, as they propose more implementable solutions for lexical grammatical transformations than the first two forms.

Another important consideration is that some features of human language appear to be of universal character, for example, every language has nouns and verbs. Even the differences of human languages often have systemic structure [1]. Syntactically languages are most different in the basic word order of verbs, subjects, and objects in declarative clauses. English is an SVO language, while Russian has a comparatively flexible word order. The syntactic distinction is connected with a semantic distinction in the way languages map underlying cognitive structures onto language patterns, which should be envisaged in MT implementations [2]. Besides, there exist syntactic constructions specific of a given language (such as, for example, English constructions with existential "there" and "it" as formal subjects). Sometimes, a word may have translation to a word of another part-of-speech in the target language, a word combination, or even a clause, as the English i*mplementable* is best translated into Russian as *kotoryi vozmozhno realizovat*' (*which can be implemented*). To overcome these differences the categorial and functional features of the two languages were considered, and structures of the input were made conformed to the rules of the target language by applying contrastive linguistic knowledge for implementation of the transfer model. A suitable formalism is indispensable for an algorithmic presentation of the established language transfer rules, and the language of Cognitive Transfer Structures (CTS) was developed based on rational mechanisms for language structures generation and feature unification.

The application of statistical models has considerably advanced the area of machine translation since the last decade of the previous century, however now new ideas and methods appear aimed at creating systems that efficiently combine symbolic and statistical approaches comprising different models. Both the paradigms move towards each other: more and more linguistics is being introduced into stochastic models of machine translation, and the rule-based systems include statistics into their linguistic rule systems. The procedures of analysis and translation are enhanced by the statistical data, which are taken into consideration by the "translation engine" for disambiguation of language structures. The paper is also focused on discovering the ways of the two research paradigms combination, namely, introducing statistical methods into the rule-based systems of machine translation and employment of the methods and presentations capturing human language intuition in statistical translation models with the view of enhancing the existing language processing technologies.

In statistical machine translation (SMT) the task of translating from one natural language into another is treated as a machine learning problem. This means that via training on a very large number of hand-made translation samples the SMT algorithms master the rules of translation automatically.

2 Establishment of cross-language matches and inter-structural synonymy

Segmentation and unification of utterances in the course of translation is a major task for human professional interpreters. They would even say that syntax is "interpreter's enemy". The selectivity of languages as to the choice of specific characteristics of description of one and the same situation results in numerous distinctions, and one of the most crucial of them is the degree of particularity in conveying a referential situation. Therefore, a situation which in one

language is described by means of one specific feature, in another language may require two or more characteristics. Thus, in many cases the English language is more economical (about thirty percent, according to the reports of simultaneous interpreters) [3,4] in expressing a thought than Russian. A very good illustration of this phenomenon is attributive word combinations of the "stone wall" type which when being translated into Russian in many cases require numerous additions. On the other hand, Russian input in some cases may result in an expanded English translation.

In practice the technique applied to overcome this problem is utterance segmentation which consists in sectioning a source Russian sentence into two or more utterances in the resulting English sentence.

Another important rule is the least possible change of word order. But this inflicts other unavoidable transformations, and not all of them are implementable within the framework of machine translation. For example, the general rule for interpreters: a Russian noun which appears at the very beginning of a sentence and has the form of an oblique case, i.e. indirect object standing at the beginning of a Russian sentence, should be transformed into the subject of an English sentence notwithstanding its initial syntactic role

e.g. *Na vstreche dogovorilis'…(At the meeting agreed…)*

should be translated as -

The meeting reached an agreement…

This transformation performed in the course of human simultaneous interpretation appears to be unattainable to a machine translator at the present state of the art. The requirement of denotational equivalence involves numerous lexical grammatical shifts which cause transformations of the semantic structure of an utterance [3,4]. Another regular semantic shift, that of substituting a predicate of action by the predicate of state.

e.g. *He is a member of the college team. (A predicate of state).*

On igraet v studencheskoi komande. (He plays in the students' team. A predicate of action).

Moreover, the existence of such shifts within the real text corpora inflicts complications for one more computational linguistics problem, that is text alignment, which in some cases may appear even intractable.

The following SI techniques appeared to be of use for MT design in the course of our development.

(1) Full translation of lexical grammatical forms is applied when these forms completely correspond to each other both in the source and the target languages as to their form, function and meaning.

(2) Null translation is applied when a grammatical form exists in the source and target languages but is used differently for explicating a certain referential situation.

(3) Partial translation is used when one and the same grammatical form has several content functions which differ in the source and target languages.

(4) Functional substitution is employed when the functions and meanings of grammatical forms in the source and target languages differ. In that case the source form can be substituted by a form of another type in the target language on the basis of their functional identity.

(5) Assimilation is a device applied for translating grammatical forms constituting compound structure, and the combinability features of these forms differ in the source and target languages.

(6) Conversion is used for substituting a form of one category by a form of another category, and is conditioned by the combinability rules difference in the source and target languages.

(7) Antonyms employment is used for eliminating a conflict between lexical and grammatical combinability of language units in the source and target languages.

Thus it is obvious that the search for equivalence should be carried out starting with the establishment of semantic equivalence of patterns notwithstanding their structural dissimilarity. Pattern-matching approach for the English – Russian transfer was assumed, and the segmentation of structures of the source language was performed on the basis of functional transfer fields which were established via contrastive study of the two languages.

The transformations in focus comprise the following statistically important cases:

- Nominalization;
- Passivization;
- Adjectival – Adverbial structures transformations;
- Subject – Object transformations;
- Indirect Object transformation into Subject;

e.g. *Ser'oznymi raznoglasiyami byla otmechena vstrecha storon – Serious disagreements arose during the meeting of the sides*

Direct Object transformation into Subject;

Prepositional phrase transformation into Subject:

Na vstreche dogovorilis' – The meeting reached the conclusion.

3 Cross-level focus

The machine translation technique employed presupposes three stages: analysis, transfer and generation. The stage of analysis results in parse representing the structure of the input sentences. Transfer is a bridge between the parse structure of the source language and the input to the generation procedure for the target language. At this stage the transformation is performed of one parse tree (applicable for the source language presentation) into another tree (presenting the target language). Thus syntactic transformations imply the mapping of one tree structure to another.

It is very important that a parse for MT differs from parses required for other purposes. Thus the grammar formalisms developed for a unilingual situation (phrase stucture rules systems for the English language) [5] would give an untransferable parse in many crucial situations. For example, just one English phrase structure rule for simple sentence would suffice for grammar parse without translation, but for the English – Russian transfer a multiple structure of possible parses is required depending on the specific finite verbal form constituting the sentence. And to overcome this, an accurate scheme for all the particular verbal form cases should be designed.

The segmentation of phrase patterns used for the input language parse was carried out with the consideration of semantics to be reproduced via the target language means. Both the most important universals such as enumeration, comparison, modality patterns, etc., and less general structures

were singled out and assigned corresponding target language equivalents.

Consider an example of a phrase structure conveying the modal meaning of obligation: "…the task to be carried out…". In other words, the meaning of this phrase can be rendered as "…the task that should be carried out…". The Infinitive phrase in the Engish language gives the regular way of expressive means compression without the loss of semantic value. A literary translation in Russian requires the second way of presenting the same idea of obligation. However in this specific case a "reduced" translation variant is also possible which consists in the introduction of the subordinate conjunction "chtoby" – "so that", between the noun and the modifying Infinitive. The parse rule would look like:

NP(to) -> NP VPto

And the generation rule would be presented as:

NP(to) -> NP Punct.{comma} Conj.(chtoby) VPto

Special attention is required for the problem of passive constructions transfer. As in the phrase "was considered". The rules for simultaneous translation (which in many cases is similar to the real time machine translation performance and can be a source of compromise decisions for phrase structure design) requires the transformation of the English Subject into the Direct Object (Russian, Accusative Case) standing in the first position in a sentence and the passive verbal form would produce an impersonal verbal form in Russian. However such transformation proved to be of considerable danger to the whole sentence structure and might cause an unpredictable generation result. Hence, for many cases a more clumsy, though robust method of a passive construction generation was accepted: the one similar to the English "be + Past Participle":

V(aux_ppt) -> V(aux) PPt

For any MT design scheme there exist major concerns such as verb subcategorization presentations, discontinuous structures treatment, phrasal units adjustment. In the English-Russian transfer these concerns are aggravated by the high productivity of the English phrasal verbs (and other units) and their derivatives.

a) An example of a phrase structure rule for the verb subcategorization:

V/np_p_inf --> Vinf NP Pt Vto_inf

get the sample down to observe

b) An example of a discontinuous structure:

Or watch the things, you gave your life to, broken

c) Phrasal units:

later on; over there; what a {good idea}.

Our approach employs both phrase structure rules and vocabulary-driven methods for dealing with these problems.

4 Generalized Cognitive Structures Underlying Transferable Syntaxemes

Actually the process of transfer goes across the functional – categorial values of language units. A language structure which can be subjected to transfer has to be semantically complete from the point of view of its function. The cases of categorial shifts, in particular, when the technique of conversion is employed, require special treatment: the categorial shift of a syntax unit is determined by the functional role of this unit in a sentence (e.g. noun as a modifier --> adjective). Only by creating the centaur concepts.. 'constituency-dependency', 'linearity-nonlinearity', 'form-function', etc. can we get a reasonably clear picture of linguistic reality [6].

The starting idea for the language structures segmentation strategy was the notion of functional semantic fields. The system of grammar units, classes and categories with generalized content supplementary to the content of lexical units, together with the rules of their functioning, is a system which in the end serves for transmission of generalized categories and structures of mental content which lie the foundation of utterance sense, and constitute the basis of language grammar formation [7].

As it was exhibited in [8] language coding technique is to a great extent determined by the deep semantic structure, and of considerable advantage is such a presentation method which takes for the starting point the semantic level, and particular semantic units are confronted with the coding devices expressing them. The approach of functional semantics concords in many aspects with the categorial grammar. The system of sentence members (functional roles) is being modified, but its essence is preserved in the new facts qualification via the traditional categories [9].

The transferability of phrase structures is conditioned by the choice of language units in the source and target languages belonging to the same functional transfer fields (FTF), notwithstanding the difference or coincidence of their traditional categorial values. A set of basic FTF was singled out and language patterns employed for conveying the functional meanings of interest were examined.

• Nomination and Relativity FTF: language structures performing the nominative functions (including the sentential units) comprise this field.

• Primary Predication FTF (non-inverted) bearing the Tense – Aspect – Voice features; this field mainly includes all possible complexes of finite verbal forms and tensed verbal phrase structures.

• Secondary Predication FTF bearing the features of verbal modifiers for the Primary Predication FTF. Included here are the non-finite verbal forms and constructions, and subordinate clauses comprising the finite verbal forms. All these are united by the functional meanings they convey, e.g. qualification, circumstance, taxis (ordering of actions), etc.

• Modality and Mood FTF: language means expressing modality, subjunctivity and conditionality are included here. Here the transfer goes across the regular grammatical forms and lexical means (modal verbs and word combinations) including phrasal units.

• Connectivity FTF: included here are lexical – syntactic means employed for concatenation of similar syntactic groups and subordination of syntactic structures.

• Attributiveness FTF: adjectives and adjectival phrases in all possible forms and degrees comprise the semantic backbone of this field; included here are also other nominal modifiers, such as nominative language units and structures (stone wall constructions, prepositional genitives – of – phrases), and other dispersed language means which are isofunctional to the backbone units.

• Metrics and Parameters FTF: this field comprises language means for presenting entities in terms of parameters and values, measures, numerical information.

• Partition FTF: included in this field are language units and phrase structures conveying partition and quantification (e.g. some of, part of, each of, etc.).

• Orientation FTF: this field comprises language means for rendering the meaning of space orientation (both static, and dynamic).

• Determination FTF: a very specific field which comprises the units and structures that perform the function of determiner (e.g. the Article, which is a good example for grammar – lexical transfer from English into Russian, since in Russian there exist no such grammatical category; demonstrative pronouns, etc.).

• Existentiality FTF: language means based on be-group constructions and synonymous structures (e.g. sentential units with existential there and it as a subject: there is…; there exists…; etc.).

• Negation FTF: lexical – syntactic structures conveying negation (e.g. nowhere to be seen, etc.).

• Reflexivity FTF: this field is of specific character since the transfer of reflexivity meaning goes across lexical - syntactic – morphological levels.

• Emphasis – Interrogation FTF: language means comprising this field are grouped together since they employ grammar inversion in English.

• Dispersion FTF: individual language structures specific for a given language are included here; these are presented as phrasal templates which include constant and variable elements.

The set of functional meanings together with their categorial embodiments serve the source of constraints for the unification mechanism in the formal presentation of our grammar. The formalism developed employs feature-based parse, and head-feature inheritance for phrase structures which are singled out on the basis of functional identity in the source and target languages. To implement the feature-valued inheritance sometimes broader contexts are taken.

5 The Existing Formalisms Influence

Our implementation formalism was developed taking into account the apparata of phrase structure and unification grammars: Head-Driven Phrase Structure Grammars (HPSG) [10], Generalized Phrase Structure Grammars (GPSG) [11], Revised Generalized Phrase Structure Grammars (RGPSG) [12]. Categorial and Dependency [13] grammars were also considered. Important for us was the strict lexicalism principle of the HPSG, i.e. word structure and phrase structure are governed by independent principles. Roles are determined by verbal valences; utterances are a blend of categorial meanings and role meanings and their structural projections which are specific for every particular language.

The technique of categorization, i.e. generation of a backbone grammar of atomic categories from distinct sets of feature bindings was first suggested by [14]. It was used for building a shift-reduce table for the Alvey grammar [5], but there it was necessary to subsume each category into its most general unifying category, so reducing the overall number of categories.

Generally, for the Russian language, dependency grammars have been applied. And phrase structure approach seemed to be less applicable here. Hence, of particular interest for us was the study and comparison of both the formal approaches, so that practical algorithmic solutions could be worked out.

A certain key was suggested in the coexisting systems of Immediate Dominance (ID) rules and phrase structure (PS) rules in ANLT [5] based on a variant of GPSG. The ID rules encode unordered dependency relations and further are subjected to linearization to be applied for the parse. GPSG may be thought of as a grammar for generating a context-free grammar. The generation process begins with immediate dominance (ID) rules which are context-free productions with unordered right-hand sides. An important feature of ID rules is that nonterminals in the rules are not atomic symbols (e.g. NP). Rather, GPSG nonterminals are sets of [feature, feature-value] pairs. For example, [N +] is a [feature, feature-value] pair, and the set {[N +], [V -], [BAR 2]} is the GPSG representation of a noun phrase. Next, metarules apply to the ID rules, resulting in an enlarged set of ID rules. In the RGPSG the finite closure problem is used to determine the cost of metarule application. Principles of universal feature instantiation (UFI) apply to the resulting enlarged set of ID rules, defining a set of phrase structure trees of depth one (local trees). One principle of UFI is the head feature convention which ensures that phrases are projected from lexical heads. Finally, linear precedence statements are applied to the instantiated local trees. The final result is a set of ordered local trees, and these are equivalent to the context-free productions in a context-free grammar. The process of assigning structural descriptions to utterances consists of two steps in GPSG: the projection of ID rules to local trees and the derivation of utterances from nonterminals, using the local trees.

In GPSG there are three category-valued features : SLASH which marks the path between a gap and its filler with the category of the filler; AGR which marks the path between an argument and the functor that syntactically agrees with it (between the subject and matrix verb, for example); and WH which marks the path between a wh-word and the minimal clause that contains it with the morphological type of the wh-word. In RGPSG the revision is unit feature closure: to limit category-valued features to containing only 0-level categories., i.e. 0-level categories do not contain any category-valued features. GPSG's ID/LP format models the head parameter and some free word order facts. The HPSG formalism is based on phrase structure rules, but dominance relations are implemented via head elements. Phrasal types are also treated in terms of multiple inheritance hierarchies that allow generalizations about diverse construction types to be factored into various cross-cutting dimensions.

In fact, each non-linear dependency rule is an encoded potential for actualization of a set of possible linear phrase structures. Therefore, we assumed a more computationally practical approach (to our knowledge, never used before in a bilingual situation), that of feature-valued head-driven phrase structures for both English and Russian.

6 Implementation Techniques

In conclusion, it should be noted that this article describes the experience in creation of modern multilingual machine translation systems – the systems of phraseological translation. The extensive application of means of automation allowed to essentially reduce expenditures of human labour in the process of creation of this system, and therefore, to reduce the creation cost of such systems. The primary purpose in introducing feature structures and unification has been to provide a way to express syntactic constraints that would be difficult to express using the mechanisms of context-free grammars alone. The next step was to design a way to integrate feature structures and unification operations into the specification of a grammar.

This was performed by augmenting the rules of the hybrid grammar comprising context-free and context-dependent rules with attachments that specify feature structures for the constituents of the rules, along with appropriate unification operations that express constraints on those constituents. These attachments were used to associate complex feature structures with lexical items and instances of grammatical categories; to lead the composition of feature structures to larger grammatical constituents based on the feature structures of their component parts; to lay compatibility constraints between specific parts of grammatical constructions. Functional meanings of units were encoded in functional tags for phrase structures, and the feature-value types were determined by functional – categorial semantics, for example:

[Feature,EnumVerb]; [Category,bePlus]; [Category,toPlusInfinitive]; [Feature,verbModal] [Feature,verbComplex];], etc.

Such major problems as reference resolution and long distance dependencies are also treated within the framework of feature-valued phrase structures.

The demand for practicality, quick implementation and low computational cost were of prime concern.

The principle of effort economy was observed: if something could be represented by weaker means, no stronger instruments were applied. We acquired the “blow-up” strategy for language structures simulation, which means that the most functionally relevant subsystems were introduced first, and then these were expanded, specifying structures being gradually included.

A constraint-based formalism comprising some features of the HPSG was developed. The formalism provides representation mechanisms for the fine-grained information about number and person, agreement, subcategorization, as well as semantics for syntactic representations. The system of rules based on this formalism can be called the Cognitive Transfer Grammar and consists of transferable phrase structures together with the transfer rules which are combined within the same pattern. Such patterns, or Cognitive Transfer Structures (CTS), are constitutional components of the declarative [15] syntactical processor module and encode both linear precedence and dependency relations within phrase structures. The CTS presentation was worked out under a certain influence of the content-based attribute structuring approach assumed in dataflow basic components [16].

The syntax of a CTS can be given as follows:

CTS -> CTS<identifier> CTS<token> <Input Phrase Structure & Feature-Value Set> <Head-Driven Transfer Scheme> <Generation Feature-Value Set & Phrase Structure >

The Cognitive Transfer Grammar provides translation of phrase structures within one CTS,

e.g. him to come -> chtoby on prishel.

A CTS rule is either a context-free or context-dependent production, and the derivational process may alternate between an AND-transition and OR-transition, these two devices introduce lexical and structural ambiguity, which is a central property of natural languages. Disambiguation techniques are based on learning methods [17].

“Abstract” structures are avoided wherever possible, in favor of constituent structures. Linguistic information is hierarchically organized in such a way as to predict the impossibility of certain kinds of linguistic phenomena. The head features inheritance is widely used. Needed feature structures are copied from children to their parents, which turns out to be a specific instance of a much more general phenomenon in constraint-based grammars. Specifically, the features for most grammatical categories are copied from one of the children to the parent. The child that provides the features is called the head of the phrase, and the features copied are referred to as head features.

In our approach the direct encoding of possible subcategorization features is made via a verbal CTS. Since the verbs can subcategorize for quite complex frames composed of many different phrasal types, we first established a list of possible phrasal types that can make up these frames, e.g. VPto “I want to know”; VPing “He contemplates using them”; Sto “feel themselves to be relatively happy”. Each verb allows many different subcategorization frames.

If compared with the existing phrasal subcategorization frames [18,19], in our system the emphasis is laid on functional motivation

7 Rule set for training data: cognitive semantic approach

In conclusion, it should be noted that this article describes the experience in creation of modern multilingual machine translation systems – the systems of phraseological translation. The extensive application of means of automation allowed to essentially reduce expenditures of human labour in the process of creation of this system, and therefore, to reduce the creation cost of such systems. In contrast to the approaches on the basis of “translation memory” that provide the increase of a machine translation system language competence by accumulating the previously translated text fragments and mainly based on regular expressions, Cognitive Transfer Grammar is intended for the realization of the mechanism of structural memory, which simulates language competence of an adult learner (“Adult Learning Memory”). Thus, structural memory comprises the following components:

1) The initial basic collection of grammar rules represented in the formalized form (CTG);

2) The mechanisms of expansion and refinement of the system of rules, implemented by means of the methods of machine learning on parallel texts.

Our studies are based on the concepts of the functional approach, which we have used for the multilingual situation. With the development of the linguistic processor, which ensures English - Russian and Russian - English transfer, we introduced the concept of functional transfer fields (FTF) that served the basis for the segmentation of language structures for the solution of machine translation problems. The basic idea of FTF consists in the adoption of the hypothesis about the fact that at the basis of grammatical structures there lie the cognitive structures (mental frames); a functional transfer field reflects the interaction of elements from different language levels.

The basic design unit of the spaces of cognitive transfer is a transfeme.

Definition. Transfeme is a unit of cognitive transfer the, i.e. a semantic element embodied in a translatable semantically relevant language segment taken in the unity of its categorial and functional characteristics, that establishes the semantic correspondence between the language structures, which belong to different language levels and systems. The types of transfemes are determined by the rank of transfemes.

We distinguish the following ranks of transfemes:

• rank 1: lexemes as structural signs, i.e., a word, considered as a categorical - functional unit without taking into account the specific lexical value of this word;

• rank 2: a word combination, i.e., the syntactic structure, which consists of two and more syntactically connected words, but never a complete sentence (clause);

• rank 3: a clausal unit, i.e., dependent (subordinate) clause;

• rank 4: a sentence (either a simple sentence or the main clause of a complex sentence);

• rank 5: a scattered structure, i.e., a word group, which is characterized by a syntactic and semantic unity, but is discontinuous, i.e., between the members of the group there appear other language objects, which are not the members of this group;

• rank 0: the morphological units, which are not independent words, but which form a part of a lexeme of a source language, and in the language of transfer can be expressed by a clause and the units of other ranks, for example: the suffixes – ible, - able which are synonymous to the construction "which can be", e.g. extensible – which can be extended.

he key idea of our linguistic framework is cognitive cross-linguistic study of what can be called configurational semantics, i.e. the systemic study of the language mechanisms of patterns production, and what meanings are conveyed by the established types of configurations. We explore the sets of meanings fixed in grammar systems of the languages under study. Our studies are focused on the types of meanings outside the scope of lexical semantics, and we consider the lexical semantics when the meanings which we denote as configurational, have expression at the lexical level. The importance of this aspect is connected with the fact that natural languages are selective as to the specific structures they employ to represent the referential situation. However, it is always possible to establish configurations which perform the same function across different languages (i.e. isofunctional structures). The parse aimed at transfer procedures requires a semantic grammar and cannot be efficiently implemented through a combination of monolingual grammars.

In the Cognitive Transfer Grammar (CTG), the functional meanings of language structures are determined by the categorial values of head elements. The probability characteristics are introduced into the rules of the unification grammar as weights assigned to the parse trees.

For the alignment of parallel texts the transfemes are given as the rewrite rules in which the left part is a nonterminal symbol, and the right part are the aligned pairs of chains of terminal and nonterminal symbols which belong to the source and target languages :

$T \rightarrow \langle \rho, \alpha, \sim \rangle$,

where T is a nonterminal symbol, ρ and α are chains on terminal and nonterminal symbols which belong to the Russian and English languages, and ~ is a symbol of correspondence between the nonterminal symbols occuring in ρ and the nonterminal symbols occuring in α. In the course of parallel texts alignment on the basis of the CTG the derivation process begins with a pair of the linked starting symbols and , then at each step the linked nonterminal symbols are rewritten pairwise with the use of the two components of a single rule.

For automatic extraction of the rules on the basis of CTG from parallel texts these texts should be previously aligned by sentences and words. The extracted rules base on the wordwise alignments in such a way that at first the the starting phrase pairs are identified with the use of the same criterion as the majority of statistical models of translation employing the phrase-based approach, which means that there should be at least one word inside a phrase in one language aligned with some word inside a phrase in another language, but no word inside a phrase in one language can be aligned with any word outside its pair phrase in another language.

8 Acknowledgements

The work presented in the paper was supported by the Russian Foundation for Basic Research, grant 11-06-00476-a.

9 Conclusion

The urgency of the new hybrid methods of language objects presentation is caused by the demand for the optimal combination of advantages of the two research paradigms: logical linguistic modelling employing the designed rules and stochastic approach based on machine learning [19-21]. This development is of special importance for the tasks of structural analysis and computer modelling of the full text scientific and patent documents. The work with patent documents requires the introduction of specific features of patent texts: such as employment of certain language constructions, the syntax of patent formulae, the extensive use of templates, domain-oriented lexicons. The Intertext base comprises a collection of scientific and patent texts in the Russian and English languages from the areas of Computer Science, Social Monitoring, Chemical Technology and other areas. One of the latest developments is connected with implementing the natural language web service for the multilingual search and analysis of financial information.

The objectives of the prospective research and development efforts consist in the inclusion of parallel texts and language

processing features for the French, German and Italian languages, and evolving the Intertext into a multilingual knowledge base. Our focus on configurations provides high portability to the language processing software designed under these principles: we can operate with a lexicon which has only standard linguistic information including morphological characteristics, part of speech information and the indication of transitivity for verbs. The approach taken would be important in further development of educational programs for computer science and computational linguistics courses. Educational relevance of the methods discussed in the paper lies in deeper understanding of uniform cognitive mechanisms employed in particular language embodiments of semantic structures.

10 References

[1] Comrie, B. Language Universals and Linguistic Typology. Basil Blackwell, Oxford. Second edition. 1989.

[2] Nirenburg, S., Carbonell, J., Tomita, M., and Goodman, K. Machine Translation: A Knowledge-based Approach. Morgan Kaufmann. 1992

[3] Visson, L. From Russian Into English: An Introduction to Simultaneous Interpretation. Ann Arbor, Michigan: Ardis, 1991.

[4] Visson, L. Syntactical Problems for the Russian-English Interpreter. No Uncertain Terms, FBIS, vol. 4, N 2, 1989, 2-8.

[5] Grover, C., Carroll, J. and Briscoe, T. The Alvey Natural Language Tools Grammar (4-th Release). Technical Report, 1993, Computer Laboratory, University of Cambridge, 1993.

[6] Shaumyan, S. A Semiotic Theory of Language. Indiana University Press, 1987. .

[7] Bondarko A.V. Printsipy Funktsional'noi Grammatiki I Voprosy Aspektologhii. Moskwa, URSS, 2001 / Functional Grammar Principles and Aspectology Questions. Moscow, URSS, 2001 (In Russian).

[8] Kibrik A.E. Ocherki po Obstchim I Prikladnym Voprosam Yazykoznaniya. Moskwa, URSS, 2002. / Studies in General and Applied Linguistics Issues. Second Edition. Moscow, URSS, 2001 (In Russian).

[9] Zolotova G.A. Kommunikativnye Aspekty Russkogo Sintaksisa. Moskwa, URSS, 2001/ Communicative Principles of the Russian Syntax. Moscow, URSS, 2001 (In Russian).

[10] Pollard, C. and Sag, I.A. Head-Driven Phrase Structure Grammar. University of Chicago Press, Chicago, 1994.

[11] Gazdar, G., Klein, E., Pullum, G. and Sag,I. Generalized Phrase Structure Grammar. Oxford: Basil Blackwell, 1985.

[12] Ristard, E.S. Computational complexity of current GPSG theory. Proceedings of the 24-th Annual Meeting of the Association for Computational Linguistics. Columbia University, New York: Asociation for Computational Linguistics. 1986, pp. 30-39.

[13] Mel'cuk, I.A. Dependency Syntax: Theory and Practice, State University of New York Press, 1988.

[14] Gazdar, G. and Mellish, C. Natural Language Processing in Prolog. Wokingam, UK: Addison-Wesley, 1989.

[15] Kozerenko E.B. Portable Language Egineering Solutions for Multilingual Processors // Proceedings of the International Conference on Artificial Intelligence IC-AI'02// CSREA Press, 2002, pp. 447-453

[16] Arlazarov V.L., Emelyanov N.E. Document Processing Systems. Basic Components. /Data Flow Management. Ed. Prof. Arlazarov V.L., Prof. Emelyanov N.E. – Moscow: Editorial URSS, 2002. (in Russian).

[17] Missiourev A. Hand-printed Character Recognition by Neural Networks.// Proceedings of the 5-th German-Russian Workshop on Pattern Recognition and Image Understanding (GRWS98), 1999.

[18] Baker, C.F., Fillmore, C.J., and Lowe, J.B. The Berkeley FrameNet project. In COLING/ACL-98, pp. 86-90, 1998.

[19] B*rown P.F., Cocke J., Della Pietra S.A., Della Pietra V.J., Jelinek F., Lafferty J.D., Mercer R.L., Roossin P.S.* A statistical approach to machine translation // Computational Linguistics, 1990. Vol. 16. P. 79–85.

[20] *Och F.J., Ney H.* A Systematic Comparison of Various Statistical Alignment Models // Computational Linguistics, 2003. Vol. 29. No. 1. P. 19–51.

[21] *Koehn P. and Hoang H.* Factored translation models // Proceedings of the 2007Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007. P. 868–876.