Features and Categories Design for the English-Russian Transfer Model

Elena Kozerenko
Institute for Informatics Problems of the Russian Academy of Sciences,
44 Corpus 2, Vavilova Str., 119333, Moscow, Russia
elenakozerenko@yahoo.com

Abstract. The paper focuses on the role of features for the implementation of the transfer-based machine translation systems. The semantic content of syntactic structures is established via the contrastive study of the English and Russian language systems and parallel texts analysis. The notion of cognitive transfer is employed which means that a language unit or structure can be singled out for transfer when there exists at least one language unit or structure with a similar meaning in the target language The approach taken is aimed at providing computational tractability and portability of linguistic presentation solutions for various language engineering purposes.

Keywords: Machine translation, syntactic structures, features, categories, cognitive transfer.

1 Introduction

The rapid development of language processing systems within the statistical frameworks has revealed the limitations of purely statistical methods in machine translation projects, and at the same time stimulated the new approaches to linguistic rule systems design making them adjusted to be used together with the modern statistical methods. The rule system described in this paper builds on the awareness of the fact that the meaning of a structure in a source language may shift to another category in the language of translation. This awareness is very important for obtaining reliable statistical information from parallel texts corpora to be further used in statistical machine translation algorithms. Otherwise the existing stochastic methods for language processing bring a lot of excessive inconsistent rules which still require filtering and hand editing.

Generally, major efforts connected with natural language modeling lay emphasis at lexical semantics presentations and less attention is paid to the semantics of structures and establishment of functional similarity of language patterns as a core problem in multilingual systems design. The studies presented in this paper focus on the semantics of language structures, namely, the interaction of categorial and functional meanings for subsequent language engineering design of feature-value structures. The proposed methods of dealing with syntactic synonymy of structures (isofunctionality) and structural (syntactic) polysemy provide an essential linguistic foundation for learning mechanisms.

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 123-138

Received 14/10/07
Accepted 07/12/07
Final Version 15/01/08

An important consideration in our work 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 can be translated by a word of another part of speech in the target language, by a word combination, or even by a clause. The parse aimed at transfer procedures requires a semantic grammar and cannot be efficiently implemented through a combination of monolingual grammars.

This paper focuses on the role of features for the implementation of the transfer-based machine translation systems. The approach taken is aimed at computational tractability and portability of the language presentation solutions for various language engineering purposes. An attempt is made to build a generalization over the feature-value sets of the English and Russian languages to introduce the functional semantic motivation into the categories of language structures. Our theoretical conclusions result from linguistic research of paradigmatic and syntagmatic properties of the languages under study and machine translation developments.

We argue that language engineering presentations should be well-grounded linguistically, hence a detailed and profound theoretical study of language features is indispensable for further language simulation. On the other hand, the computational linguistic models shouldn't be overloaded with detailed rules. We operate with a set of 21 basic features for the Russian language and a set of 18 basic features for the English language. The maximal difference of feature-value structures is observed when comparing the nominal features (a system of cases in the Russian language and the system of cases in the English language). The meanings of "genitivity", "dativity", "instrumentality", etc. have to be represented via non-morhpological means – syntactic configurations, i.e. word order and prepositional phrases. We avoid introducing any abstract categories, unless they naturally "emerge" from the basic features and their configurations. So, we are guided by a "reasonable sufficiency" principle in working out the system of features and values for the Russian and English languages which accords with the minimalist program.

Another dramatic mismatch between the languages is connected with the behavior of the English Nonfinite Verbal Forms, such as Gerunds and Participles. Their double nature directly affects the transfer into the Russian language: we have to choose either verbal or nominal match in the Russian language for the Gerund, and either adjectival or adverbial interpretation for the Participle. Again we have to introduce the syntactic level means, since there are no direct morphological correspondencies in Russian. We construct a presentation system employing the basic morphological features and their combinations to capture the meanings of the higher level units and structures – syntactic and semantic features. Categorial feature structures serve the objective for introducing functional semantics: category is the potential for function (functions).

The presentation mechanism chosen resembles that of HPSG with its attention both to constituency and dependency relations [3]. It is an attempt to build a phrase structure generative system for the Russian language having the objective of multilingual transfer and establishing the synonymy of language structures in a multilingual situation. The set of functional meanings together with their categorial embodiments serves 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. Phrase structures are singled out on the basis of their functional identity in the source and target languages which accords with the approach of [4]. Important evidence of robustness of statistical methods employment in rule-based parse is given in [5].

The parse and transfer procedures accuracy ranges from 34.75 to 73.43 in our model.

2 Language Structures Transferability

To face the problems of language structures transferability for machine translation (MT), it is of great help 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 from the Web, getting acquainted with subject domain information, etc., a rough translation output without post editing can be quite acceptable. 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.

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) 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.

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 Cognitive Transfer Fields which were established via contrastive study of the two languages [8].

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.

The absence of full coincidence between the English and Russian language constructions can be found when studying the comparative frequency of the parts of speech use, especially in scientific and technical texts. In general the style of scientific discourse is characterized by a greater rate of "nominativity", i.e. the use of nouns than the other styles. And the comparative study of translations shows that this tendency is considerably stronger in the Russian language where the verbs of the source English texts are frequently substituted by nouns. Our studies show that the Russian text is approximately by 35% more nominal than the English text. Consider the following examples of verbal-nominal transformations in the English-Russian translations.

These considerations are important for building translation systems employing machine learning methods.

3 Semantic Match Establishing Principles

Studying the categorial and functional meanings of language structures we have established that important tools realizing these meanings are ways of configuring phrase structures, i.e. linearization patterns: possible linear sequences of language objects (of units and structures).

Semiotic linguistics [9] calls these ways of configuring *structural signs* and also introduces the concept of *superposition* of functions, presuming that every language object has its primary function, and shifts of meanings which occur in speech (i.e. language in action) are superposition of secondary and other functions onto the primary one.

Our research is focused on revealing all possible types of structural signs which convey similar meanings, i.e. establishment of syntactic synonymy.

A classical example of syntactic synonymy is the means of expressing case meanings, e.g. morphological in the Russian language (through cases endings) and analytical in English - by means of prepositions and the order of words.

Hence, our problem is to reveal all types of structural signs and compose them into a uniform system of semantic syntactical representations for a language processor.

Superposition of functions is a very useful tool for construction of functional semantic representations of linguistic units and structures.

The concepts of primary and secondary functions of language signs enable us to create the new representations using traditional categories (as, for example, the Verbal_Noun = Verb + Noun).

The method applied for creating a system of rules for functional transfer in machine translation is described in [10]. The establishment of structures equivalence on the basis of functional semantics proved to be useful for developing the syntactic parse and transfer rules module for the English – Russian machine translation [8]. Generally, major efforts connected with natural language modeling lay emphasis at lexical semantics presentations and less attention is paid to the semantics of structures and establishment of functional similarity of language patterns as a core problem in multilingual systems design.

Our interpretation techniques employ the segmentation of structures carried out on the basis of the functional transfer principle. The principal criterion for including a language structure into a field is the possibility to convey the same functional meaning by another structure of the field, i.e. the interchangeability of language structures. To establish whether the structures and units are equal or not, we need some general equivalent against which the language phenomena would be matched. In Contrastive Linguistics the notion of *tertium comparationis* is widely employed to denote this general equivalent, and the approach based on the principle "from the meaning to the form" focusing on Functional Syntax would yield the necessary basis for equivalence search.

What differs our approach is the attention to the semantics of configurations, i.e. the study of the way languages tend to *arrange structures* in order to convey certain meanings. And we focus on the linear patterns of the languages under study, since we assume that linearization is not a random process but it is determined by the cognitive mechanisms of speech production and the way they manifest themselves in syntactic potentials of a given language. The primary object of our contrastive language study was to establish what particular language meanings are represented in the categorial-functional systems of the English and Russian languages. Categorial values embody the syntactic potentials of language units, i.e. their predictable behavior as syntactic structures (syntaxemes). Thus, as it was demonstrated in [9-11], Category is the potential for Function, and multiple categorial values inflict multiple syntactic functions. However, when we analyze language in action, i.e. the utterances of written or sounding speech, the function of a particular language structure determines which of the potentials is implemented in this utterance, hence which of the possible categorial values of a polysemous syntactic structure is to be assigned to the parse result.

4 Cognitive Transfer Structures

Our observation is that the cognitive linguistic 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 [9].

The starting idea for the language structures segmentation strategy was the notion of functional semantic fields [7] in the Russian language. 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 lay the foundation of utterance sense, and constitute the basis of language grammar formation.

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

Primary Predication CTF (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 CTF bearing the features of verbal modifiers for the Primary Predication CTF. 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.

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

Modality and Mood CTF: 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 CTF: included here are lexical – syntactic means employed for concatenation of similar syntactic groups and subordination of syntactic structures.

Attributiveness CTF: 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 CTF: this field comprises language means for presenting entities in terms of parameters and values, measures, numerical information.

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

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

Determination CTF: 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 CTF: 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 CTF: lexical – syntactic structures conveying negation (e.g. nowhere to be seen, etc.).

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

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

Dispersion CTF: individual language structures specific for a given language are included here; these are presented as phrasal templates which include constant and variable elements. To implement the feature-valued inheritance sometimes broader contexts are taken.

A constraint-based formalism of Cognitive Transfer Grammar (CTG) was developed and implemented in the English-Russian machine translation system [8]. It comprised 222 transferable phrase structures together with the transfer rules combined within the same pattern. 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 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 syntactical processor module of the machine translation system and encode both linear precedence and dependency relations within phrase structures.

The initial variant syntax of a CTS was designed 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. This was the approach employed in the first version of the English-Russian machine translation system which provided one variant of translation to a CTS. At present a Multivariarnt Cognitive Transfer Grammar (MCTG) has been designed which envisages several translations for each CTS. It comprises about 350 transferable phrase structures together with the multivariant transfer rules.

Consider, for example, the functional meaning of Possessiveness, which belongs to the CTF of Attributiveness in the following phrases:

Peter's house; the house of Peter

These phrases have the same meaning, that could be presented by the following semantic network:

Owner ← Having → Had Thing.

However, we see our main objective not in creation of an abstract semantic meta language, but in the careful research of all possible kinds of configurations of language categories, used by natural languages for expression of functional meanings.

To determine the categorial and functional values of language structures we employ the technique of Compositional Categories. This technique consists in the superposition of categorial values of language objects to envisage the possibility of multivariant translations. Thus the Gerund is categorized as "VerbNounIng", the

Infinitive in the Subject function is categorized as "toPlusInfinitiveSubj", the Russian Adverbial Participle ("Deeprichastie") and the the English Participle Active in its adverbial function are categirized as "ParticipleAdv", finite verb forms are referred to as "VerbFinit", other categiries are also used, as for example, the following:

{[Category: VerbNounIng]: *asking questions*};

{[Category: toPlusInfinitiveSubj]: *She is known to be a skilled typist* };

{[Category: toPlusInfinitiveObj]: *We feel them to be sensitive readers*}.

5 Polysemy and Ambiguity of Syntactic Structures

By syntactic polysemy we mean the immediate realization of more than one categorial meaning within the head element of a language structure. The polysemous structures display variable manifestation of their categorial features depending on the functional role in the sentence. Consider such language phenomena as the Gerund, the Participle and the Infinitive.

The Gerund comprises the features of both the Verb and the Noun, which affects the translation strategy when the appropriate means are to be chosen for representation of the English Gerund via the Russian language forms. The structures similar in category to the English Gerund are the Russian Verbal Nouns denoting "Activity", e.g. *singing* → *penie, reading* → *chtenie*, and both the English Gerund, and the Russian Verbal Noun allow direct object arguments if derived from transitive verbs. However, the direct transfer of the Gerund into the Russian Verbal Noun is the least probable translation variant of the three possible transfer schemes:

The Gerund (Eng) → Clause with the Finite Verb form (Rus)

The Gerund (Eng) → Clause with the Infinitive (Rus)

The Gerund (Eng) → Verbal Noun (Rus).

This fact can be accounted for by the mechanisms employed in the Russian language for configuring sentential structures and is to be envisaged in the machine translation engine.

Consider the other most productive polysemous language structures which comprise more than one categorial meaning:

The Participle → Verb + Adjective

The Infinitive → Verb + Noun

Nominal Phrase as the Nominal Modifier → Noun + Adjective

Verbal Phrase as the Verbal Modifier → Verb + Adverb.

Thus we introduce the notion "polysemous syntactic structure" to determine the set of possible transfer schemes for a given language structure. When a polysemous structure is assigned specific categorial attributes realized in this structure, the possible and preferable transfer schemes become predictable for the given structure.

The predominant categorial meaning of a polysemous syntactic structure (or syntaxeme) is determined by the syntactic function realized at a given moment. Thus the transfer scheme for a "*stone wall*" construction will be as follows:

Noun1 + Noun2 [Eng.] → Adjective + Noun2 [Rus]

The weight for this transformation will be higher than for the transformation:

Noun1 + Noun2 [Eng] → Noun2 + Noun1 (Genitive) [Rus]

if the dictionary contains an Adjective as one of the possible translation equivalents for Noun1, that is the case when the dictionary is composed by various methods including acquisition of lexical units from parallel texts.

Judging by the function we establish the Cognitive Transfer Fields (CTF) within which the translation procedure will be carried out CTF support the possible paraphrasing variants and envisage the synonymous ways of conveying the same functional meaning across languages.

Of special interest is the situation of the categorial shift in translating a syntactic pattern. The category of a syntactic pattern, i.e. a phrase structure, is determined by the category of the head word of this phrase structure. Thus, when transfer employs conversion, and the category of the head word shifts to another category, the whole structure is assigned the feature of the new category. Thus a Nominal modifier of a Nominal Phrase becomes an Adjective in translation; a Verbal unit acting as a Verbal modifier becomes an Adverbial clause containing the Finite Verbal form. The latter case accords with the SUG principle of the Verb being the Sentence Nucleus [9,11].

To illustrate the mechanism of polysemous structures transfer we take the Secondary Predication CTF and the Attributiveness CTF.

The Secondary Predication CTF bearing the features of verbal modifiers for the Primary Predication structures (the non-inverted Finite Verb forms and tensed verbal phrase structures bearing the Tense – Aspect – Voice features) includes the nonfinite 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.

The following schemes of transfer into Russian are applicable to the phrase:

Feeling surprised seemed permanent.

"Gerund + Participle II + Finite Verbal Phrase" → " Sentence " →

"Nominative Clause + Finite Verbal Phrase" (1)

OR

"Verbal Noun Phrase + Finite Verbal Phrase" (2)

The Participle in postposition to a Nominal Phrase most frequently would be transferred into a Russian Clause:

The material processed satisfied all the requirements.

"Nominal Phrase + Participle II + Finite Verbal Phrase" → " Sentence " →

"Nominal Phrase + Qualifying Clause + Finite Verbal Phrase" (1)

OR

"Nominal Phrase + Participle II + Finite Verbal Phrase" (2)

Attributiveness CTF: 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 nominal language units and structures (*stone wall* constructions, prepositional genitives – *of* –phrases), and other dispersed language means which are isofunctional to the backbone units.

Consider the phrases of the kind: "*a woman of means*", "*a man of talent*". Possible contexts might be as follows:

She was a woman of means.

He was a man of talent.

The multivariant transfer would comprise the following Russian phrase structures:

(1) with the Genitive construction;

(2) with the Qualifying Clause;
(3) with the Preposition "*s*" (Russian):
"Nominal Phrase1 + of +Nominal Phrase2" →
"Nominal Phrase2 + Nominal Phrase1 Genitive"
Or
"Nominal Phrase1 + Qualifying Clause"
Or
"Nominal Phrase1 + Prep "*s*" + Nominal Phrase2 Instrumental".

The last variant would mean in Russian "*a woman with means*", "*a man with talent*".

We took into account the computational cost of the rule system which led us to a certain minimalism: we avoided introduction of abstract categories in rule design (having in mind the imperative known as Ockham's Razor: the notion that when presented with a choice of axioms or laws, or explanations, it is wise to choose the one that is the *simplest*). All the functional meanings were presented as feature – value structures based on traditional language categories.

We find it important to differentiate between polysemous and ambiguous syntactic structures. A polysemous structure implies possible realizations of meanings which are compatible within one language structure and can be transferred to the structures of another language which are isofunctional to the source language structure. An ambiguous syntactic structure presupposes alternative ways of interpretation, the meanings being incompatible within one language structure, thus we deal with ambiguity when we try to discern some Finite and Nonfinite verbal forms:

Gerund / Present Participle;
Infinitive / Present Simple;
Past Participle / Past Simple.

Ambiguous structures can be misleading to the parsing procedures and subsequent machine translation, as for example, the "garden path" is a well-known language phenomenon which may give incorrect parse at the early stage of analysis, that could be corrected only at the final stage:

The cars passed by the vessel drowned.

The possible interpretations for the sentence can be as follows:

The cars which were passed via the vessel drowned (the correct variant).
The cars which passed the vessel drowned.

However, the phrase"*The new control system updated continuously displayed robust performance*" was analyzed and translated correctly by all the tested modern MT systems which comprise learning mechanisms within their framework. This fact can be explained by the presence of the broader context.

6 Disambiguation Techniques: Rule-Based and Machine Learning Methods

The impact of differentiation between syntactic polysemy versus syntactic ambiguity consists in the following implementation decisions. An ambiguous structure is analyzed in alternative manner: each possible parse and transfer variant is presented

as a separate rule, and constraints are introduced into the rule structure. A polysemous structure is assigned a multiple transfer scheme within one rule.

The mechanism of computational (contextual) reframing (CR) is being designed for treatment of the two major bottlenecks: syntactic derivation history (for words in a secondary, tertiary, etc. syntactic function) and syntactic polysemy of structures. Reframing models the use of the same structural unit in different structural and/or lexical contexts, which results in the difference of the meanings of this unit. The presentations for the syntactic module rest on the basis of traditional word categories. Contextual correlations associated with each function of a structural unit are established via stochastic data obtained from corpora study.

Since parse procedures sometimes may result in more than one possible structure, the rules and lexical entries are supplied with the probabilistic augmentations which serve for syntactic ambiguity resolution. The multivariant rules that envisage the variants of transfer for polysemous structures and separate alternative rules for ambiguous structures have been worked out. They comprise the most probable methods of the language structures transfer, as for example, the Infinitive constructions in the function of the Adverbial Modifier of Goal/Consequence:

Hydrogen and oxygen unite to form water.

The scheme of the multivariant English-Russian transfer of the construction *to form water* will be as follows:

[Category: VerbInf] → {*to form water* }

OR {[Category:ParticipleAdv]; {*образуя воду – forming water*}

[Category: VerbFinit]; {*образуют воду – form water*}

[Category: VerbNounIng]} {*с образованием воды – with formation of water*}

In the course of sentence analysis and parent nodes formation the resulting structures will be marked-up by the compositional categories which provide the appropriate transformations of language structures for transfer.

As natural language generates an infinite number of sequences, learning mechanisms are incorporated into the parse engine: information about unfamiliar words and structures can be inferred from the context. The data on which the inference can be founded is accumulated by learning on parallel texts: a supervised algorithm is trained on a set of correct answers to the learning data, so that the induced model may result in more accurate decisions.

The lexical model employs the concise lexicon entries presenting categorial, morphological and combinatorial information supplied with the statistical data for each lexical item characterizing its distribution.

We studied the existing results in the field of human cognitive mechanisms of language learning, as well as machine learning methods: there is substantial evidence that the way children learn their first language may be understood as information compression [12,13]; the Optimality theory states the importance of grammatical architecture with the strict prioritization or ranking, rather than any scheme of numerical weighting.

Of particular interest for us was the study and comparison of various formal approaches [3,5,14-28], so that practical algorithmic solutions could be worked out, we adhere the strict lexicalism principle of the HPSG [3], i.e. word structure and phrase structure are governed by independent principles.

It is a well-known fact that the underlying tree representation plays a crucial role for the performance of a probabilistic model of grammar [14]. Probabilistic developments of the unification grammar formalism are given in [15,17,18]. A new probabilistic model for Combinatory Categorial Grammar is presented in [16].

The phenomenon of syntactic polysemy determines the possible multiple transfer scheme for a given language pattern. We develop the system of multivariant transfer rules - Multivariant Cognitive Transfer Grammar (MCTG).

The probability values for syntactic analysis variants can be obtained either on the basis of corpus information, or from linguistic expert knowledge. In the latter case we deal with reliable information fixed in grammar systems of languages distilled by the centuries of human language practice.

The values of probabilities for every possible parse variant (i.e. the expansion of a nonterminal node) are calculated on the basis of frequencies of occurrence of each analysis variant in the existing text corpora with syntactic mark-up (treebanks). The calculation is made of the number of times (N), when some variant of expansion of a node ($\alpha \to \beta$) is used with subsequent normalization:

$$P(\alpha \to \beta \mid \alpha) = \frac{N(\alpha \to \beta)}{\sum_{\gamma} N(\alpha \to \beta)} = \frac{N(\alpha \to \beta)}{N(\alpha)} \tag{0.1}$$

The probability of the full parse of a sentence is calculated with the account of categorial information for each head vertex of every node. Let n be a syntactic category of some node n, and h (n) is the head vertex of the node n, m (n) is a mother node for the node n, hence, we will calculate the probability p(r(n)|n, h(n)), for this we transform the expression (1.2) in such a way, that every rule becomes conditioned by its head vertex:

$$P(T,S) = \prod_{n \in T} p(r(n) \mid n, h(n)) \times p(h(n) \mid n, h(m(n))) \tag{0.2}$$

Since the ambiguity of some syntactic structure (a node) is understood as an opportunity of realization of more than one categorial value in the head vertex of this structure, the probability of the full parse of a sentence containing ambiguous structures (i.e. nodes, subtrees) will be calculated with the account of the probabilities of the categorial values realized in the head vertices of these structures (nodes).

In our grammar system the functional values of languages structures are determined by the categorial values of head vertices. The probabilities are introduced into the rules of the unification grammar CTG as the weights, assigned to parse trees. Ambiguous and polysemous syntactic structures are modeled by the Multivariant Cognitive Transfer Grammar Structures (MCTGS).

The syntax of a MCTG structure (MCTGS)can be presented in the following way:

MCTGS → MCTGS<identifier> MCTGS<weight> MCTGS<mark-up> <Input Phrase with Feature-Value Structue> <Head-Driven Scheme of Transfer> <Generated Phrase with Feature-Value Structure 1> <weight 1> <Generated Phrase with Feature-Value Structure 2> <weight 2> ...<Generated Phrase with Feature-Value Structure N> <weight N>

The MCTG approach provides the mechanisms for modelling the transfer procedures for polysemous and ambiguous syntactic structures and it can be extended for a greater number of languages.

7 Language Engineering Environment

Our current project INTERTEXT is aimed at creation of systemic presentations of functionally motivated semantic syntactic structures. The above stated methods are being employed for design and development of a language engineering environment comprising the research linguistic knowledge base Phrasenet and the features for multivariant parse and transfer of language structures. It is a linguistic resource with semantic grouping of phrase structure patterns provided with the links to isosemic structures at all language levels for the whole set of languages included into the linguistic base. The categorial systems of a subset of natural languages (English, Russian and some other European languages) and functional roles of language units in a sentence have been explored and the core set of transferable language phrase structures has been established on the basis of generalized cognitive entities manifested in the grammar systems under study. The structures are segmented on the semantic principle of functional transferability, i.e. these structures should be "translatable".

Our linguistic simulation efforts are aimed at capturing the cross-level synonymy of language means inside a system of one natural language and interlingual semantic configurational matches. This emphasis on the practical human translation experience gives the reliable foundation for statistical studies of parallel text corpora and automated rule extraction in further studies.

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, and no detailed semantic mark-up is required for lexical entries.

The Phrasenet linguistic knowledge base comprises the following components:

- a parallel texts database: the texts are segmented into the functionally relevant structures that are semantically aligned; - a structural parse editor (under development at present) which displays the parse and transfer schemes for indicated text segments; - a multilingual functional treebank; - a functional semantic vocabulary of structural configurations arranged on the basis of the Cognitive Transfer principle.

Special Attention is given to the phenomena of syntactic polysemy and ambiguity.

8 Conclusions

Our analysis and development experiments result in understanding the efficiency of what might be called the "exteriorization of meaning", i.e. accumulation of relevant data concerning the functional-categorial features of possible structural contexts and/or specific lexical contexts that help to disambiguate the parsed structures of the

source language and decide what particular meaning of a language structure is realized in the given text segment. Rather than invent a sophisticated antropocentric heuristics for the rule-based disambiguation techniques via traditional linguistic presentations, we need to design a synthetic mechanism comprising the core rule set and reliable learning methods.

The rule set applicable both for the transfer procedures and for acquiring new linguistic data by corpora study should envisage the phenomena of syntactic polysemy and ambiguity of structures. The solution employed in our project is based on the Cognitive Transfer Structures (CTS) approach grouping isofunctional language structures, and the Multivariant Cognitive Transfer Grammar (MCTG) comprising the rules which state the multiple equivalent structures in the source and target languages. The MCTG linguistic rule set is being augmented by Probabilistic Functional Tree Substitution features. Since the nodes of the MCTG have functional articulation, the trees and subtrees of the possible parses also have functional character, i.e. are tagged by functional values.

Our further research and development efforts are connected with the refinement of the existing presentations, inclusion of specific lexical-based rules into the grammar system, and excessive corpora-based experiments for extending the mechanisms of multiple transfer. Our attention at present is directed at the improvement of statistical mechanisms incorporated in the system of parse and transfer rules for the English-Russian language pair and designing a model which includes other languages (Italian and French).

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. Pollard, C. and Sag, I.A. Head-Driven Phrase Structure Grammar. University of Chicago Press, Chicago, 1994.
4. Mustajoki, A. Functional Syntax as the Basis for Contrastive Languae Study. Studia Slavica Finlandesia. Finnish Contributions to the 13th International Congress of Slavists, Ljubljana, August, 15-21, 2003, pp.100-127 (In Russian).
5. Briscoe, E.J. "An introduction to tag sequence grammars and the RASP system parser". Technical Report N662, University of Cambridge, March, 2006.
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. Kozerenko, E.B. Cognitive Approach to Language Structure Segmentation for Machine Translation Algorithms // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 23-26, 2003, Las Vegas, USA.// CSREA Press, pp. 49-55, 2003.
9. Shaumyan, S. Categorial Grammar and Semiotic Universal Grammar. In Proceedings of The International Conference on Artificial Intelligence, IC-AI'03, Las Vegas, Nevada, CSREA Press, 2003.

10. Kozerenko, E.B., Shaumyan, S. Discourse Projections of Semiotic Universal Grammar // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 27-30, 2005, Las Vegas, USA.// CSREA Press, pp. 3-9, 2005.
11. Shaumyan, S. Intrinsic Anomalies in Meaning and Sound and Their Implications for Linguistic Theory and Techniques of Linguistic Analysis. // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 27-30, 2005, Las Vegas, USA.// CSREA Press, pp. 10-17, 2005.
12. Chater, N. Reconciling simplicity and likelihood principles in perceptual organization Psychological Review 103 (3), pp. 566-581, 1996.
13. Chater, N. The search for simplicity: a fundamental cognitive principle? The Quarterly Journal of Experimental Psychology. 52 (2), 273-302, 1999.
14. Johnson, M. PCFG Models of Linguistic Tree Representations. Computational Linguistics, 24(4), pp. 613-632, 1998.
15. Osborne, M. and Briscoe, T. Learning Stochastic Categorial Grammars. In T.M. Ellison, editor, Proceedings of CoNLL97: Computational Natural Language Learning, pp. 80-87, Somerset, NJ, 1998.
16. Hockenmaier, Julia. Data and Models for Statistical Parsing with Combinatory Categorial Grammar. Ph.D. thesis, University of Edinburgh , 2003.
17. Brown, P.F., J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer & P.S. Roossin. A statistical approach to machine translation. Computational Linguistics 16, pp. 79-85, 1990.
18. Habash, Nizar and Bonnie Dorr. Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation. AMTA-2002. Tiburon, California, USA, 2002.
19. Dorr, Bonnie and Nizar Habash. Interlingua Approximation: A Generation-Heavy Approach. AMTA-2002 Interlingua Reliability Workshop. Tiburon, California, USA, 2002.
20. Apresjan, Jurij, Igor Boguslavsky, Leonid Iomdin, Alexandre Lazursku, Vladimir Sannikov, and Leonid Tsinman. "ETAP-2: the linguistics of a machine translation system".META, 37(1):97–112, 1992.
21. Boguslavsky, I., Chardin, I., Grigorieva, S., Grigoriev, N., Iomdin, L., Kreidlin, L. and Frid, N. "Development of a dependency treebank for Russian and its possible applications in NLP". In Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Gran Canaria, Spain, 2002.
22. Kakkonen T. "Dependency treebanks: methods, annotation schemes and tools". Proceedings of the 15th NODALIDA conference, Joensuu, 94-104, 2005.
23. Copestake, A., Lascarides, A., & Flickinger, D. "An algebra for semantic construction in constraint-based grammars". In Proceedings of the 39th Meeting of the Association for ComputationalLinguistics. Toulouse, France, 2001.
24. Flickinger, D. "On building a more efficient grammar by exploiting types". Natural Language Engineering, 6 (1) (Special Issue on Efficient Processing with HPSG), 15 – 28, 2000.
26. Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. "Building a large annotated corpus of English. The Penn Treebank". Computational Linguistics, 19, 313 – 330, 1993.
26. Oepen, S., Callahan, E., Flickinger, D., & Manning, C. D. "LinGO Redwoods. A rich and dynamic treebank for HPSG". In LREC workshop on parser evaluation. Las Palmas, Spain, 2002.
27 McCarthy, D., and Carroll J. "Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences" Computational Linguistics, vol. 29.4, 639-654, 2003.
28. Watson, R., J. Carroll and E. Briscoe "Efficient extraction of grammatical relations", Proceedings of the 9-th International Workshop on Parsing Technologies (IWPT'05), Vancouver, Ca., 2005.