Multilingual Systems: Grammar Acquisition by Machine Learning

© Elena Kozerenko

Institute for Informatics Problems of the Russian Academy of Sciences
kozerenko@mail.ru

Abstract

Multilingual systems design is discussed from the point of view of machine learning methods applicability for grammar acquisition. The key idea is developing a synergistic approach combining semantic grammar rules with the machine learning mechanisms of grammar rules extraction from text corpora. This work is the further development of the methods and the rule system presented in [16], which were employed in machine translation system "Cognitive Translator".

1 Synthetic Approach to Multilingual Systems Design

This paper highlights some pressing issues of natural language syntax and semantics which are of prime importance to the tasks of multilingual systems design and implementation. The approach proposed consists in synergistic employment of linguistic rule-based methods presenting the core grammar for the tasks of machine translation and knowledge management, and machine learning techniques for new grammar rules acquisition and disambiguation of structures.

The following fundamental features of the given work differentiate it from other projects:

- the emphasis on the semantics of grammar, i.e. the study of language configurations basing on the Functional Transfer Fields (FTF) [16] and their projections into particular language structures; this will result in developing a computational variant of a multilingual semantic grammar (MSG), and the semantic grammar applicable for language learning by students;
- MSG is a projection of the Multivariant Cognitive Transfer Grammar (MCTG) designed in our project for the English-Russian correspondencies onto a wider range of languages including German, French and Italian;
- MSG development is dominant to the lexical semantic studies which are conducted on the basis of the existing English-Russian computational vocabulary obtained from parallel text corpora;
- The semantic grammar is used for the establishment of regular correlates between structures (configurations) and lexical units, i.e. cross-level correspondences;
- the construction of the systemic cross-lingual presentation of phrase structures conveying the similar meanings in the languages under study and employing it as the core of the rule system for the algorithms of syntactic-semantic transfer in machine translation, multilingual knowledge management and language learning systems;
- the synthetic approach enables the system to generalize the rules and avoid the overgeneration of rules thus resulting in translation accuracy improvement.

The texts of business and scientific discourse have been studied, and specific structures are extracted and serve the source for further rule set development. Experiments will be made with the texts of corporative documents and patents.

2 Machine Learning Techniques in Natural Language Processing

The research in the NLP area is substantially stimulated by the fact that the market for MT grows mature in 2002 - 2004, and more corporations realize that the implementation of a customized MT solution can give a great advantage to the company and make them lead in the competition for today's multilingual customers.

The aim of machine learning is to infer automatically a model for some domain on the basis of the given data from the domain, thus a system learning syntactic rules would be supplied with a set of phrase structure rules to be used for training. Recently more attention has been paid to the construction of N-grams capturing sophisticated presentations of syntactic and semantic structures: using long-distance triggers instead of local N-grams [23,20], applying variable-length N-grams [19], including semantic information to the N-grams, e.g. semantic word associations based on the latent semantic indexing [14].

The learning algorithms can be of two types: unsupervised and supervised. An unsupervised algorithm has to induce a model capable of generalizing to new data it hasn't been given before, and does this purely from the data. A supervised algorithm is trained on a set of correct answers to the learning data, so that the induced model would result in more accurate decisions. At present we employ the latter approach in our work using Bayesian methods operating on the phrase structure weights obtained from the text corpora. This model will be developed into the probabilistic functional tree substitution grammar (PFTSG) mechanism operating with the system of multivariant cognitive transfer rules. Machine learning is largely founded on the stochastic research paradigm rooted in the development of speech recognition algorithms, character recognition, spelling correction. An essential probabilistic framework employs the Bayes Rule and the noisy channel model which play an important role in many problems, in particular, part-of-speech tagging and probabilistic parsing. The algorithm which is essentially employed within this architecture is known as dynamic programming algorithm.

The recent decade has witnessed a considerable progress of natural language processing (NLP) techniques based on machine learning. The appearance of large parallel texts corpora promoted the statistical methods of NLP which now augment the scheme of the principal existing approaches to machine translation (MT) design – direct translation, transfer and interlingua-based methods. A statistical machine translation was first introduced by [5,6].

The starting point for any natural language processing system is tagging module design. Different stochastic taggers appeared in the 1980s. The idea shared by all stochastic taggers consists in choosing the most likely tag for a given word.

One of the most popular probabilistic taggers is the Hidden Markov Model (or HMM tagger) - for a given sentence or word sequence, HMM taggers pick the tag sequence that maximizes the following formula: P(word| tag) * P(tag|previous n tags).

An approach to machine learning based on rules and stochastic tagging is known as the Transformation-Based Learning (TBL). TBL is a supervised learning technique, and employs a pre-tagged training corpus.

For probabilistic parsing stochastic grammars are applied. A Probabilistic Context-Free Grammar (PCFG) is a 5-tuple G = (N,T,P,S,D), where N is a set of non-terminal symbols, T is a set of terminal symbols, P is a set of productions of the form A -> *b*, where A is a non-terminal symbol, *b* is a string of symbols, S is a designated start symbol, D is a function assigning probabilities to each rule in P.

Probabilistic tree substitution gramar (PTSG) : the definition is the same as PCFG but rather than a set of rules, we have a set of tree fragments of arbitrary depth whose top and interior nodes are nonterminals and whose leaf nodes are terminals or nonterminals, and the probability function assigns probabilities to these fragments. PTSGs are thus a generalization of PCFGs, and are stochastically more powerful , because one can give particular probabilities to fragments - or even whole parses - which cannot be generated as a multiplication of rule probabilities in a PCFG.

We considered both TBL and PTSG when working out the way how to develop the stochastic learning apparatus for our formalism of Multivariant Cognitive Transfer Grammar (MCTG). The accepted approach in the initial variant was the use of Bayesian methods operating on the phrase structure weights.

The development of speech-to-speech translation systems [15,10] substantially stimulated the research in the field of MT. The existing computational resources provided for today's MT systems allow to accumulate and recall previously corrected translations (Translation Memory and Example-based Machine Translation) [27,7]. A model for machine translation based on an aligned text corpus is example-based machine translation, which means that the example-based translation employs the closest match in aligned corpora as a template for translation. The descriptions of example-based MT systems are given in [18,24].

The evidence of the latest research and development projects shows that machine learning methods alone are unlikely to yield the finely tuned language processing decisions.

However the use of combined techniques brings the increase of "meaningful" performance of language processing systems in different aspects, thus it was exhibited in [25] that introduction of linguistic (morphological) parse rules into the search engine considerably enforces the precision of search results.

New solutions and combinations of methods are being set forward. It is possible to say that at the present moment the systems solely based on translation memory are giving way to the systems which comprise several complementary techniques: though the systems based on the principles of Translation memory produce understandable translations, they still lack grammatical accuracy.

3 The State of the Art in Present Day Machine Translation

When considering the performance of the systems we relied rather on subjective evaluation of translations correctness than any tools giving numeric expression of translation quality. The task to develop such tools is of great complicity, though at present some interesting methods have been set forward [13].

The projects were analyzed taking into account underlying models employed, the degree of particularity in semantic presentations, methods used for language parse and generation, performance and functionality aspired for, and the degree of human participation in detailed description of language phenomena. Basing on these criteria the projects fall into the four groups.

The first group is constituted by the projects which lay emphasis on reducing the human intervention into the process of language acquisition and disambiguation. Statistical MT systems based on training data sets on parallel corpora, where rules are extracted automatically from texts, are developing methods that automatically identify the most relevant contextual features for determining the sense of any ambiguous word or word combination. The advantages of this approach consist in fully or to a great extent automating the process of linguistic knowledge base formation. However, the bottleneck emerges on the other side: the rules automatically constructed lack accuracy and are overgenerated, which requires post-editing of rules and special efforts and techniques for exclusion of invalid and superfluous rules. Irrelevant rules are eliminated basing on statistics, but the major diffficulty is rule generalization: automatic rule extraction gives many variants for one and the same rule, and this situation is to be handled by human linguists. Thus in many cases a predesigned rule set would be indispensable.

The noisy channel model of statistical machine translation and example-based machine translation are the two most widely used statistical MT models.

In the following works [1,2,31,33] attempts are made to replace the statistical word-for-word approach with a statistical transfer approach.

The most successful stochastic language models have been implemented on the basis of finite-state presentations, such as N- grams or hidden Markov models. However, finite-state models cannot represent hierarchical structures found in natural language.

Since we lay emphasis on the synergy of methods, of special interest for us were the projects and implemented systems based on the translation memory approach or including example-based and learning-based techniques as one of their features, for they are giving the complementary solutions for the problems which cannot be treated efficiently by logical approach alone, namely, computationally relevant presentations of contexts. We studied the MT projects available in the Internet, tested their performance (when possible) and compared them with traditional rule-based MT systems. The best-known developments of this group are as follows.

SDLX of the SDL International employs translation memory (TM) and supports all languages having the Latin, Arabic and Hebrew alphabets [34]. TM is the basic feature of the products launched by the company in 2003: SDLinsight and SDLX Translation Suite 2003.

DIPLOMAT machine translation system [35]. Rapid deployment is achieved through the use of Pangloss's example-based machine translation (EBMT) and transfer-based MT, within the Multi-Engine MT architecture, which uses a statistical target language model to help select between competing translations. This technology also makes the system's user-driven incremental improvement possible.

"eAccela BizLingo" is web-based English-Japanese and Japanese-English translation server software that enables all employees on a company intranet to translate documents, email, and web pages quickly and easily. eAccela BizLingo adds value to an enterprise by accelerating knowledge sharing and facilitating information access between international business partners. Furthermore, when used in conjunction with eAccela BizSearch, it can serve as a crosslingual search system for web servers, file servers, or groupware containing multiple documents or emails in both English and Japanese; the developer is Fujitsu Software Corporation [36]. The more specific and domain-oriented a system is, the more robust performance can be expected from it. Thus a high degree of precision show such systems as, for example, ENGSPAN, SPANAM which were used in the Pan American Health Organization.

“Business English” developed by the company Lingualec Sprachtechnologien and aimed at translation of business documentation from English into German and vice versa; the system also employs translation memory mechanism, its database contains more than 25000 sentences. LogoVista E to J Pro, winner of the Apple Japan Product Excellence Award, is a powerful professional system for large-scale technical translation. It includes the comprehensive LogoVista Dictionary, customizable User Dictionaries, and interactive features for refining the translation [37]. Hence, major shortcomings of machine translation tend to be reduced in three ways: by narrowing the problem scope of the texts to be translated and tuning to a subject area; by employment of shallow approaches which give results relevant for informational purposes; by creating program instruments for human translators.

The second group is constituted by the projects which aim at creating a cognitive model with a strong semantic presentation formalism [30,8,21,32]. The prime importance is given to lexical semantic presentations encoded in vocabulary entries. Often semantic descriptions are closely tied with lexical units instantiating structures, and structural behavior of language patterns to be analyzed in real text is predicted on the basis of semantic and syntactic expectations included into vocabulary entries. This approach requires apparata for fine-grained semantic presentations and highly qualified manual linguistic performance to introduce these descriptions into computational lexicons. These works enhance general understanding of human language operation. However, abounding fine-grained information results in a greater number of rules, and the number of rules is often critical since it may lead to considerable computational costs. The systems implemented on this principle can be called lexicon-driven developments.

The projects aimed at multilingual performance tend to assume the interlingua approach. Thus the KANT project, part of the Center for Machine Translation (CMT) at Carnegie Mellon University (CMU), was founded in 1989 for the research and development of large-scale, practical translation systems for technical documentation. KANT uses a controlled vocabulary and grammar for each source language, and explicit yet focused semantic models for each technical domain to achieve very high accuracy in translation. Designed for multilingual document production, KANT has been applied to the domains of electric power utility management, heavy equipment technical documentation, medical records, car manuals, and TV captions. The KANT analyzer uses a lexicon and a unification grammar to produce a set of lexical functional grammar style f-structures. Some grammatical information may also be presented in the interlingua, if it is necessary for accurate translation. Natural language structures are translated into conceptual presentations of the Ontology Works fact base which is implemented as a relational database with the facility allowing semantic predicates and relations to be instantiated via a knowledge representation language called OWL [22,17].

A large-scale series of projects for Interlingual Machine Translation (NYI) and Development of a Framework for Large-Scale Translation, Tutoring, and Information Filtering (PFF/PECASE) are described in [38]. The main goal for the projects is to investigate the applicability of a lexical-based framework to large-scale natural language processing (NLP) tasks such as interlingua machine translation (MT), foreign language tutoring (FLT), and information filtering and retrieval. The projects aim to systematize the relation between syntax and semantics in lexical representations and to examine the problem lexicon construction in multilingual NLP applications. The languages involved in the study area are English, Arabic, French, Korean and Spanish.

However, the existing interlingua approaches mostly focus on such language properties which can be captured by the predicate logic and slot-filling techniques, thus giving the content for creation of conceptual presentations. No profound interlingual study has been undertaken as yet of the configurational semantics, i.e. the meanings conveyed by the organizational properties of languages – the way languages arrange the units in a structure to produce a particular semantic effect.

The MT systems marked by longevity are based on lexicons and rule sets either incorporated into the dictionary presentations or performing as separate modules, and translation engines. Thus, the developers of Systran claim that its good performance is greatly due to the translation engine. The translation engine uses recognized linguist rules and word phrasing in the translation process. This means the software will look at each sentence and recognize how a word is used in a sentence to determine the proper translation. Systran recognizes tens of thousands of set phrases. Today, 36 SYSTRAN MT language pairs are commercially available; the system also comprises subject-specific dictionaries [39], and the Babel Fish of AltaVista and SYSTRAN can be found in [40]. The system available at the Russian market is ProMT [41]; and the research project ETAP [42] employing the mechanisms of dependency grammar is the major system featuring a profound semantics-based lexicography as the core component for the Russian-English translation. At present, most of these systems develop new features including some elements of statistical methods.

The third group is comprised by the hybrid systems applying both rule-based and stochastic approaches to language modeling. Emergence of such projects is the peculiar feature of the present day trends in NLP, the strong sides of diverse approaches are aimed at attaining the goal of efficient MT and NLP systems construction. The projects vary in the particular architecture and specific models employed.

A combination of hand-made semantic descriptions and statistics was employed in Matador, a Spanish-English machine translation system, implemented following the Generation-heavy Hybrid approach to Machine Translation (GHMT). The focus of GHMT is addressing the lack of resource symmetry between source and target languages. No transfer rules or complex interlingual representations are used. Rich target language symbolic resources such as word lexical semantics, categorial variations and subcategorization frames are used to overgenerate multiple structural variations from a target-glossed syntactic dependency representation of source language sentences. This symbolic overgeneration, which accounts for possible translation divergences, is constrained by multiple statistical target language models including surface n-grams and structural n-grams [12,9].

The fourth group is constituted by the systems categorized as translation workplaces: they set forward the tasks of creating translator assistants (translation workstations and workbenches) rather than translator substitutes.

The present day tendency of comprising the rule-based and stochastic techniques within one framework leads to the emergence of projects featuring the strong sides of diverse approaches.

4 The Principal Features of the Functional Transfer Approach

As it was stated above, our approach is founded on the interpretation techniques which employ the segmentation of structures on the basis of the functional transfer principle. The search for equivalence is carried out starting with the establishment of semantic equivalence of patterns notwithstanding their structural dissimilarity. The segmentation of structures of the source language is performed on the basis of functional transfer fields (FTF) which have been established via contrastive study of the English and Russian languages [16].

- Primary Predication FTF (non-inverted) bearing the Tense – Aspect – Voice features.
- Secondary Predication FTF bearing the features of verbal modifiers for the Primary Predication FTF.
- Nomination and Relativity FTF: language structures performing the nominative functions (including the sentential units) comprise this field.
- Modality and Mood FTF: language means expressing modality, subjunctivity and conditionality are included here.
- 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.

- 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.
- 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 encountered*, 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 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. Consider, for example, the Attributiveness FTF comprising adjectives and adjectival phrases in all possible forms and degrees and other nominal modifiers. We also include here the "of" – phrases, e.g. "people of culture" can be interpretered as
a) "people who are cultured" and
b) "people who belong to the sphere of culture".
The Russian equivalents for the two meanings will be
a) "kul'turnye liudi", i.e. the Adjectival phrase; and
b) "liudi kul'tury", i.e. the Genitive construction.
The multiple correlations have been established and introduced into the system of Multivariant Cognitive Transfer Grammar rules. This system helps to be aware of possible trans-categorial shifts that occur in parallel texts, which is important when we extend the rule system in the process of learning on corpora.

Segmentation and unification of utterances in the course of translation is a major task for human professional interpreters. 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) [28,29] in expressing a thought than Russian. 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. The requirement of denotational equivalence involves numerous lexical grammatical shifts which cause transformations of the semantic structure of an utterance [28,29].

These facts should be regarded when working with parallel text corpora, otherwise, only very rough correspondences can be obtained. The interlingual equivalents will be established taking into account the following translation techniques which can be encountered in parallel texts:

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

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

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

(4) Functional substitution: the functions and meanings of grammatical forms in the source and target languages differ.

(5) Conversion: a form of one category is substituted by a form of another category, and is conditioned by the combinability rules difference in the source and target languages.

The set of FTF will be further examined and particularized for the English, Russian and German languages with the prospect of implementing a three-language model.

Rule Hierarchy Establishment is the process of grammar formalism design over the set of functional transfer structures. In our case a generative unification grammar is employed incorporating the feature-value structures into the hybrid system of context-free (and partly context-sensitive) productions. The parse and generation is performed on the basis of these rule systems. We assume a computationally practical approach of feature-valued head-driven phrase structure rules for all the languages included.

The further development of the rules set is carried out on the basis of the following principles: the ordering of the tree structures and the head features inheritance is determined by the Nucleus Law (stating that the nucleus of a configuration takes on the function of this configuration on top of its own function of the nucleus of this configuration), and the multiple transfers is supported by our typing strategy [26].

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 structure rules systems for the English language) [3,4,11] would not give a transferable 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 is being designed. Hence, a transferable grammar cannot be efficiently implemented by a mechanical composition of unilingual grammars: a semantic approach is required, and in our case, it is the employment of the functional transfer fields.

As natural language generates an infinite number of sequences, **learning mechanisms** are envisaged. The data on which the inference is founded is accumulated by learning on parallel texts. 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.

A model for machine translation based on an aligned text corpus is example-based machine translation, which means that the example-based translation employs the closest match in aligned corpora as a template for translation. The advantages of this approach consist in fully or to a great extent automating the process of linguistic knowledge base formation. However, the bottleneck emerges on the other side: the rules automatically constructed lack accuracy and are overgenerated, which requires post-editing of rules and special efforts and techniques for exclusion of invalid and superfluous rules.

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 which are being further developed into the Multivariant Cognitive Transfer Grammar.

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 [16]. 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. Syntactic structures of natural languages in many cases tend to be polysemous and ambiguous which results in multiple possible transfers from one language to another. The systemic study of polysemous syntactic structures is carried out on the basis of functional transfer principle, and cross-lingual correspondences between the English and Russian languages are encoded in the multivariant cognitive transfer rules. 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. sing*ing* -> pen*ie,* read*ing* -> chten*ie*, 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

The Gerund (Eng) -> Verbal Noun.

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 transfer field [16] within which the translation procedure will be carried out. The Functional Transfer Fields (FTF) support the possible paraphrasing variants and envisage the synonymous ways of conveying the same functional meaning across languages. To illustrate the mechanism of polysemous structures transfer we take the *Secondary Predication* FTF 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 + ParticipleII + 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 + ParticipleII + Finite Verbal Phrase"
-> " Sentence " ->
"Nominal Phrase + Qualifying Clause + Finite Verbal Phrase" (1)
Or
"Nominal Phrase + ParticipleII + Finite Verbal Phrase" (2)

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 studies presented in this paper focus on the semantics of language structures, namely, the functional meanings. The proposed methods of dealing with syntactic synonymy of structures (isofunctionality) and structural (syntactic) polysemy will provide an essential linguistic foundation for learning mechanisms.

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. 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.

The categorial systems of a subset of natural languages (English and Russian in our case) and functional roles of language units in a sentence have been explored and the core set of transferable language phrase structures has been established [16] on the basis of generalized cognitive entities manifested in the grammar systems under study. The formalism developed for presentation of syntactic structures for the English-Russian machine translation is a variant of unification grammar and comprises about 347 rules. The initial set of 222 rules has been developed into the Multivariant Cognitive Transfer Grammar (MCTG). We assume the “blow-up” strategy for language structures simulation, which means that the most functionally relevant subsystems are introduced first, then the model is gradually expanded and specifying structures included accordingly. The model will be developed into a cross-lingual semantic grammar comprising multivariant rules to envisage the mechanism of syntactic polysemy resolution.

The logical rule system is supported by statistical contextual data obtained basing on linguistic intuition of experts and from the restricted text corpora. It is associated with each particular meaning of a language structure.

5 Handling Multivariant Transfer by MCTG

The multiple correlations have been established and introduced into the system of Multivariant Cognitive Transfer Grammar rules. This system helps to be aware of possible trans-categorial shifts which occur in parallel texts which is important when we extend the rule system in the process of learning on corpora.

The constraint-based formalism of the Multivariant Cognitive Transfer Grammar consists of transferable phrase structures together with the transfer rules which are combined within the same pattern. Such patterns, or Multivariant Cognitive Transfer Structures (MCTS), serve constitutional components of the declarative syntactical processor module and encode a) linear precedence, b) dependency relations within phrase structures, c) weights for each possible transfer of a source structure. The syntax of a MCTS can be given as follows:

MCTS -> MCTS<identifier> MCTS <weight> MCTS<token> <Input Phrase Structure & Feature-Value Set> <Head-Driven Transfer Scheme> <Generation Feature-Value Set & Phrase Structure 1 ><weight 1> <Generation Feature-Value Set & Phrase Structure 2 > <weight 2> …<Generation Feature-Value Set & Phrase Structure N ><weight N>

The Multivariant Cognitive Transfer Structure provides translation of phrase structures within one bloc,

e.g. *them to develop -> chtoby oni razrabotali.*

A MCTG 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. “Abstract” structures are avoided wherever possible, in favor of constituent structures. Linguistic information is hierarchically organized in such a way as to filter out certain kinds of linguistic phenomena. The head features inheritance is widely used. Needed feature structures are copied from children to their parents, which is a specific instance of constraint-based grammars.

Our formalism of the Multivariant Cognitive Transfer Grammar (MCTG), being developed at present, actually employs the mechanism of the Probabilistic Functional Tree Substitution Grammar (PFTSG) and comprises the presentation facilities both for constituency and dependency, as well as disambiguation instrumentality.

PFTSG conforms with the Probabilistic tree substitution gramar (PTSG): PFTSG is a 5-tuple G = (N,T,P,S,D), where N is a set of non-terminal symbols, T is a set of terminal symbols, P is a set of productions of the form A -> *b*, where A is a non-terminal symbol, *b* is a string of symbols, S is a designated start symbol, D is a function assigning probabilities to each rule in P. Important here is that we have a set of tree fragments whose top and interior nodes are nonterminals representing the fuctional values of phrase structures and whose leaf nodes are terminals or nonterminals, and the probability function assigns probabilities to these fragments.

6 Implementation Techniques

The principal implementation goal was to design a way to integrate feature structures and unification operations into the specification of a grammar. This was performed by introducing 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. 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 consists of transferable phrase structures together with the transfer rules which are combined within the same pattern.

In our approach the direct encoding of possible subcategorization features is made via a verbal MCTS. 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.

7 Conclusions

We see our principal objective in developing a novel synthetic approach where language structures of several natural languages are aligned on the basis of functional meanings conveyed by a concise system of categorial values presented in cross-lingual charts, lexical and structural disambiguation is performed by means of stochastic techniques and new structures and patterns are acquired with the help of machine learning methods.

The MCTS approach provides a concise and extensible platform for cross-lingual transfer of language structures functional meanings and can be applied to a greater number of languages (especially with similar categorial feature-value structures). The problems of discontinuity, reference resolution and ambiguity are addressed by learning techniques on the basis of the employed rule system. Our further research is focused on resolution of syntactic polysemy by introducing special feature-value augmentations to the existing presentations for tracing the discontinuous structures, specifying the semantic values of particular head features, verbal subcategorization frames and numerous phrasal units adjustment.

The proposed investigation would be important in further development of educational programs for computer science and computational linguistics courses. The functional approach would introduce new semantics-based methods in language grammar studies. Educational relevance of the methods proposed lies in deeper understanding of uniform cognitive mechanisms employed in particular language embodiments of functional semantic structures. Introduction of the Multilingual Semantic Grammar approach into language teaching courses will also enhance understanding of semiotic language mechanisms operation.

References

[1] Alshawi, H., Bangalore, S., and Douglas, S. Automatic acquisition of hierarchical transduction models for machine translation. In COLING/ACL-98, Montreal, pp. 41-47. ACL. 1998.

[2] Alshawi, H, Adam L. Buchsbaum, and Fei Xia. A comparison of head transducers and transfer for a limited domain translation application. In ACL 35/EACL 8, pp. 360-365, 1997.

[3] Briscoe, T. and Carroll, J. Generalized Probabilistic LR parsing of natural language (corpora) with unification-based grammars. Computational Linguistics, 19(1), 25-59, 1993.

[4] Briscoe, T. and Carroll, J. Automatic extraction of subcategorization from corpora. In Fifth Conference on Applied Natural Language Processing, Washington, D.C., pp. 356-363. ACL, 1997.

[5] 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.

[6] Brown P.F., S.A. Della Pietra, V.J. Della Pietra and R.L. Mercer. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2), pp. 263-311, 1993.

[7] Brown, R.D. Example-based machine translation in the Pangloss system. In COLING-96, Copenhagen, pp. 169-174, 1996.

[8] Dorr, Bonnie J. and Clare Voss. A Multi-Level Approach to Interlingual MT: Defining the Interface between Representational Languages. International Journal of Expert Systems, 9(1):15-51, 1996.

[9] Dorr, Bonnie and Nizar Habash. Interlingua Approximation: A Generation-Heavy Approach. AMTA-2002 Interlingua Reliability Workshop. Tiburon, California, USA, 2002.

[10] Frederking, R., Rudnicky, A.I., and Hogan, C. Interactive speech translation in the DIPLOMAT project. In Proceedings of the ACL-97 Spoken Language Translation Workshop, Madrid, pp. 61-66. ACL, 1997.

[11] 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.

[12] Habash, Nizar and Bonnie Dorr. Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation. AMTA-2002. Tiburon, California, USA, 2002.

[13] Hovy E., King M. & Popescu-Belis A. Principles of Context-Based Machine Translation Evaluation. Machine Translation, vol. 16, pp.1-33, 2002.

[14] Jurafsky, D. and Martin, J.H. Speech and Language Processing. Prentice Hall, 2000.

[15] Kay, M., Gawron, J., and Norvig, P. Verbmobil: A Translation System for Face-to-Face Dialog. CSLI; 1992.

[16] 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.

[17] Mitamura , T. and Nyberg , E. Automatic Rewriting for Controlled Language Translation. Proceedings of the NLPRS 2001 Workshop on Automatic Paraphrasing: Theory and Application, 2001.

[18] Nagao, M. A framework of a mechanical translation between Japanese and English by analogy principle. In Alick Elithorn and Ranan B. Banerji (eds.), Artificial and Human Intelligence, pp. 173-180. Edinburgh: North-Holland, 1984.

[19] Ney, H., Essen, U., and Kneser, R. On structuring probabilistic dependencies in stochastic language modeling. Computer Speech and Language, 8, 1-38, 1994.

[20] Niesler, T.R. and Woodland, P.C. Modelling word-pair relations in a category-based language model. In IEEE ICASSP-99, pp. 795-798, IEEE, 1999.

[21] Nirenburg, S., Jaime Carbonell, Masaru Tomita, and Kenneth Goodman, eds. Machine Translation: A Knowledge-Based Approach. Morgan Kaufmann Publishers, San Mateo, CA, 1992.

[22] Nyberg, E. and Mitamura, T. The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains. In Proceedings of COLING-92 1992.

[23] Rosenfeld, R. A maximum entropy approach to adaptive statistical language modeling. Computer Speech and Language, 10, 187-228, 1996.

[24] Sato, S. CTM: An example-based translation aid system. In COLING 14, pp. 1259-1263, 1992.

[25] Segalovich I.V. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 23-26, 2003, Las Vegas, USA.// CSREA Press, pp. 273-280, 2003.

[26] 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.

[27] Sumita, E. and Iida, H. Experiments and prospects of example-based machine translation. In ACL-91, Berkeley, CA, pp. 185-192. ACL, 1991.

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

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

[30] Voss, Clare and Bonnie J. Dorr. Toward a Lexicalized Grammar for Interlinguas. Machine Translation, 10(1-2):139-180, 1995.

[31] Wang, Ye-Yi, and Alex Waibel. Modelling with structures in statistical machine translation. In ACL 36/COLING 17, pp. 1357-1363, 1998.

[32] Wilks, Yorick and Roberta Catizone: Lexical Tuning. CICLing 2002: 106-125.

[33] Wu, D. and Wong, H. Machine translation with a stochastic grammatical channel. In COLING/ACL-98, Montreal, pp. 1408-1414. ACL, 1998.

[34] Web site: SDL Machine Translation System http://www.sdlintl.com/products-home/products.htm

[35] Web site: Diplomat Machine Translation System http://www.lti.cs.cmu.edu/Research/Diplomat/

[36] Web site: BizLingo Machine Translation System http://www.fujitsu.com/services/software/translation/bizlingo/

[37] Web site: NLP Products http://www.worldlanguage.com/Products/92.htm

[38] Web site: Dorr, B.J. Large-Scale Interlingual Machine Translation (NYI) and Development of a Framework for Large-Scale Translation, Tutoring, and Information Filtering (PFF/PECASE) http://cslu.cse.ogi.edu/nsf/isgw97/reports/dorr.html

[39] Web site: SYSTRAN 4.0 http://www.translation.net/syspro.html

[40] Web site: Babelfish Machine Translation System http://babelfish.altavista.com/

[41] Web site: ProMt Machine Translation System http://www.translate.ru

[42] Web site: ETAP Machine Translation System http://proling.iitp.ru