MACHINE TRANSLATION FOR KINDRED LANGUAGES

G.Nevmerjitskaia, N.Roubashko

Faculty of Applied Mathematics and Informatics, Belarus State University, Skorina avenue, 4, Minsk, 220050, BELARUS, roubashko@fpm.bsu.unibel.by

ABSTRACT

Automatic translation from one human language to another using computers, better known as machine translation (MT), is a longstanding goal of Computer Science. In order to be able to perform such a task, the computer must "know" the two languages – synonyms for words and phrases, grammars of the two languages, and semantic or world knowledge. One way to incorporate such knowledge into a computer is to use bilingual experts to hand-craft the necessary information into the computer program. Another is to let the computer learn some of these things automatically by examining large amounts of parallel text: documents which are translations of each other.

This paper describes the problems of MT for the Byelorussian and Russian languages which are kindred languages and presents a corpus-based approach in order to receive satisfactory solution.

1. INTRODUCTION

Automatic translation between human languages (Machine Translation) is a long-term scientific dream of enormous social, political, and scientific importance. It was one of the earliest applications suggested for digital computers, but turning this dream into reality has turned out to be a much harder, and in many ways a much more interesting task than at first appeared. Nevertheless, though there remain many outstanding problems, some degree of automatic translation is now a daily reality.

The *social* or *political* importance of MT arises from the socio-political importance of translation in communities where more than one language is generally spoken. Translation is necessary for communication. Being allowed to express yourself in your own language, and to receive information that directly affects you in the same medium, seems to be an important right.

The *commercial* importance of MT is a result of related factors. First, translation itself is commercially important. Secondly, translation is expensive. Translation is a highly skilled job, requiring much more than mere knowledge of a number of languages, and in some countries at least, translators' salaries are comparable to other highly trained professionals. Moreover, delays in translation are costly. Estimates vary, but producing high quality translations of difficult material, a professional translator may average no more than about 4–6 pages of translation (perhaps 2000 words) per day, and it is quite easy for delays in translating product documentation to erode the market lead time of a new product. It has been estimated that some 40–45% of the running costs of European Community institutions are "language costs", of which translation and interpreting are the main element [1].

Scientifically, MT is interesting, because it is an obvious application and testing ground for many ideas in Computer Science, Artificial Intelligence, and Linguistics, and some of the most important developments in these fields have begun in MT.

Philosophically, MT is interesting, because it represents an attempt to automate an activity that can require the full range of human knowledge – that is, for any piece of human knowledge, it is possible to think of a context where the knowledge is required.

MT started out with the hope and expectation that most of the work of translation could be handled by a system which contained all the information we can find in a standard paper bilingual dictionary. Source language words would be replaced with their target language translational equivalents, as determined by the built-in dictionary, and where necessary the order of the words in the input sentences would be rearranged by special rules into something more characteristic of the target language. In effect, correct translations suitable for immediate use would be manufactured in two simple steps. This corresponds to the view that translation is nothing more than word substitution (determined by the dictionary) and reordering (determined by reordering rules). Especially it seems to be very easy task to translate kindred languages, e.g. Byelorussian and Russian. Translation between kindred languages is aided by resemblance of cognate word forms.

But experience shows that "good" MT cannot be produced by such delightfully simple means. As all translators know, word for word translation

doesn't produce a satisfying target language text, not even when some local reordering rules (e.g. for the position of the adjective with regard to the noun which it modifies) have been included in the system. Translating a text requires not only a good knowledge of the vocabulary of both source and target language, but also of their grammar – the system of rules which specifies which sentences are well-formed in a particular language and which are not. Additionally it requires some element of **real world knowledge** – knowledge of the nature of things in the world and how they work together – and technical knowledge of the text's subject area. Researchers certainly believe that much can be done to satisfy these requirements, but producing systems which actually do so is far from easy. Most effort in the past 10 years or so has gone into increasing the subtlety, breadth and depth of the linguistic or grammatical knowledge available to systems.

And MT between kindred languages doesn't appeare to be easy task. This paper concider some problems of such MT and offers an approach for solution of these problems.

2. Russian-Byelorussian Machine Translation

2.1. Machine Translation Engines

Traditionally, MT has been based on **direct** or **transformer** architecture engines, and this is still the architecture found in many of the more well-established commercial MT systems.

MT systems are **Linguistic Knowledge** (LK) systems. They include the two approaches that have dominated MT research over most of the past twenty years. The first is the so-called **interlingual** approach, where translation proceeds in two stages, by analyzing input sentences into some abstract and ideally language independent meaning representation, from which translations in several different languages can potentially be produced. The second is the so-called **transfer** approach, where translation proceeds in three stages, analyzing input sentences into a representation which still retains characteristics of the original, source language text. This is then input to a special component (called a transfer component) which produces a representation which has characteristics of the target (output) language, and from which a target sentence can be produced. MT system for Byelorussian and Russian is also transfer system.

The main idea behind transformer engines is that input (source language) sentences can be transformed into output (target language) sentences by carrying out the simplest possible parse, replacing source words with their target language equivalents as specified in a bilingual dictionary , and then roughly re-arranging their order to suit the rules of the target language. The overall arrangement of such an Engine is shown in fig. 1.

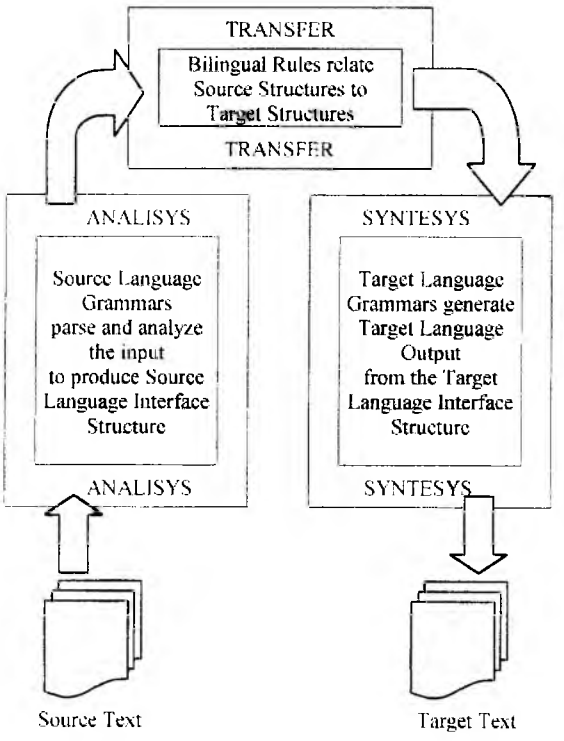

Figure 1. The components of a Transfer System

Characteristic to the performance of such a system is the fact that the engine will not be particularly troubled when faced with unusual, marginally acceptable or frankly unacceptable source language sentences; it will rarely have sufficient source language grammatical knowledge to recognise something as ungrammatical. If the grammatical structures in the input sentence are not recognised by some transforming rule, that structure will pass through to the output sentence without any re-arrangement [1].

Something similar is true for the words in the input sentence: if they are not found in the system's dictionary then they are passed through into the output and remain untranslated. As a consequence of these features this type of architecture implies that, in the worst case, the whole input sentence could survive unchanged as the output sentence. This would happen in the highly unlikely case that none of the input words are found in the bilingual dictionary and none of the input sentence grammatical structure is recognised.

With regard to the target language performance

of the system we can say that since the system has no detailed knowledge of target language grammar there is no guarantee that the transformed input sentence is actually a grammatical sentence in the target language. Although in most cases output will resemble the target language (especially the use of target language words), the result can sometimes be a completely unintelligible "word salad". In such cases one could say that the output does not belong to any known language – natural or artificial.

The typical design features of a transformer system pose some restrictions on the development of additional language modules. First, the engine will run in one direction only, for example, from Russian to Byelorusian. If the engine developer wants it to go in the other direction he more or less has to completely rewrite the transformer rules. Since the transformer rules include bilingual dictionary rules , this can mean that the Engine has to be supplied with two bilingual dictionaries, for example, Russian-Byelorusian and Byelorusian-Russian. This is rather clumsy since, apart from the differences in their directionality, the dictionaries contain much the same information. Secondly, the engine links a single pair of languages only. If the developer wants it to translate into another target language then again he/she more or less has to completely re-write the transformer rules. Even in cases where a system contains only a rather limited grammatical knowledge of the languages it involves reproducing this knowledge for the development of other language pairs means an unnecessary time loss.

Drawing these various points together, the situation of the transformer engine architecture can be summarised as follows [1]:

- It is highly **robust**. That is, the Engine does not break down or stop in an "error condition" when it encounters input which contains unknown words or unknown grammatical constructions. Robustness is clearly important for general-purpose MT.
- In the worst case it can work rather badly, being prone to produce output that is simply unacceptable in the target language ("word salad").
- The translation process involves many different rules interacting in many different ways. This makes transformer systems rather hard to understand in practice – which means that they can be hard to extend or modify.
- The transformer approach is really designed with translation in one direction , between one pair of languages in mind, it is not conducive to the development of genuinely multi-lingual systems (as opposed to mere collections of independent one-pair, one-direction engines).

2.2. General Statement of Russian-Byelorussian MT

Creating effective industrial MT systems is still a problem. But it is quite possible if a number of mathematical problems and problems of engine implementation will be resolved. MT problem can be considered as a problem of information retrieval and then the main problems are development of general system model, effective organization of the database and access methods, formalization of linguistic algorithms of text analysis and synthesis within the framework of the most suitable model of knowledge representation and construction of its optimum control structure in accordance with natural language (NL) features as an object of simulation [2].

Let's denote any NL by

$$L=(A, M, S_1, S_2) \qquad (1)$$

where:

- A – an alphabet including an «empty» character (blank), punctuation marks, etc;
- M, S_1, S_2 – sets of morphological, syntactical and semantic rules of chains formation with the help of concatenation.

Then any text T_i of NL L we can consider as a finite chain in A derivated by definite subsets $M^{(i)}$, $S_1^{(i)}$, $S_2^{(i)}$ of sets M, S_1, S_2 accordingly.

We denote by $CM(T_i)$ the description of subsets of $M^{(i)}$, $S_1^{(i)}$, $S_2^{(i)}$ for T_i and call **semantic-grammatical mark**.

Let T be a finite set of texts of one or several NL.

We say that T_i, $T_j \in T$ are in the relation R_s of identical sense if

$$CM(T_i) \equiv CM(T_j) \qquad (2)$$

It should be noticed that $CM(T_i)$ in general case includes only the description of semantic subset $S_2^{(i)}$.

As the relation R_s is an equivalence relation with a field T, it determines unique decomposition of set T on classes, each contains all $T_i \in T$ with identical $CM(T_i)$.

Consider $L^{(s)}$ as input (source) NL and $L^{(t)}$ as output (target) NL. As the task of translation is to transmit its exact sense in structural units of NL $L^{(t)}$ for $T_i^{(s)} \in L^{(s)}$, then the set-theoretic model of MT system can be described as follows:

$$M=<T,\ R_s> \quad (3)$$

where $T=T^{(s)}\cup T^{(t)}$ and for each $T_i \in T$ is indicated $CM(Ti)$.

The translation in MT system realised in accordance with model M is an identification of input text $T_i^{(s)}$ as unit $T^{(s)}$ to obtain its $CM(T_i^{(s)})$ and a selection of equivalent $T_i^{(t)} \in T^{(t)}$ in accordance with relation R_s.

As any text consists of finite set of words, phrases, sentences and discourses, a lot of new models can be obtained from model M in which any of above-stated units may be used as initial units of NL on the assumption that there are semantic-grammatical marks and sets P of the rules (algorithms) of text analysis and text synthesis. These rules are specified on higher than initial levels of structural units of NL down to a level of text.

The rules will finally realise an equivalence relation R_s for $T_i^{(t)}$ and $T_i^{(s)}$.

The selection of this or that model depends, first of all, on a certain MT problem. It should be noticed that the organization of algorithms in MT systems can be based on the so-called iconical, algorithmic and combined principles depending on ways of contrasting units of input and target languages.

Based on the said principles and general model of MT it is possible to say that the model of Russian-Byelorussian MT system is constructed on a combined principle and can be shown as:

$$M_2=<L_1^{(s)},\ L_2^{(t)};\ R,\ P> \quad (4)$$

Where:

- $L_1^{(s)}$ – a set of word forms and collocations of a source language with semantic-grammatical marks indicated for them, (realised as the systems of tags),
- $L_2^{(t)}$ – a set of word forms and collocations of a target language,
- R – a set of algorithms of word families correspondence,
- P – a set of algorithms of correction within family of words; these algorithms finally realise relation R_s.

2.3. Problems of Russian-Byelorussian MT

Let's consider some particular problems which the task of translation poses for the builder of MT systems – some of the reasons why MT is hard. They are the following [1]:

- problems of *ambiguity,*
- problems that arise from *structural* and *lexical differences* between languages;
- multiword units like idioms and collocations.

Of course, these sorts of problem are not the only reasons why MT is hard. Other problems include the sheer size of the undertaking, as indicated by the number of rules and dictionary entries that a realistic system will need, and the fact that there are many constructions whose grammar is poorly understood, in the sense that it is not clear how they should be represented, or what rules should be used to describe them. This is the case even for English, which has been extensively studied, and for which there are detailed descriptions – both traditional "descriptive" and theoretically sophisticated – some of which are written with computational usability in mind. It is an even worse problem for other languages. Moreover, even where there is a reasonable description of a phenomenon or construction, producing a description which is sufficiently precise to be used by an automatic system raises non-trivial problems.

In the best of all possible worlds (as far as most Natural Language Processing (NLP) is concerned, anyway) every word would have one and only one meaning. But, as we all know, this is not the case. When a word has more than one meaning, it is said to be lexically ambiguous. When a phrase or sentence can have more than one structure it is said to be structurally ambiguous. Ambiguity is a pervasive phenomenon in human languages. It is very hard to find words that are not at least two ways ambiguous, and sentences which are (out of context) several ways ambiguous are the rule, not the exception. This is not only problematic because some of the alternatives are unintended (i.e. represent wrong interpretations), but because ambiguities can "multiply" [1, 3].

Some of MT problems are to do with lexical differences between languages – differences in the ways in which languages seem to classify the world, and what concepts they choose to express by words. Other problems arise because different languages use different structures for the same purpose, and the same structure for different purposes. In either case, the result is that we have to complicate the translation process.

This problem seems to be not very important for Russian and Byelorussian as kindred languages which have common lexical and structural features. But there exist situations when verb requires another government, nouns have different gender in Russian and Byelorussian, or participles in Byelorussian are not translated by single word

like in Russian.

The next problem concerns the translation of idioms and collocations. Roughly speaking, idioms are expressions whose meaning cannot be completely understood from the meanings of the component parts.

The problem with idioms, in a MT context, is that it is not usually possible to translate them using the normal rules. There are exceptions which can be translated literally from one language to another and have the same meaning. But in most cases the use of normal rules in order to translate idioms will result in nonsense. Instead, one has to treat idioms as single units in translation.

There are two approaches for treatment of idioms. The first is to try to represent them as single units in the monolingual dictionaries and to construct special morphological rules to produce these representations before performing any syntactic analysis – this would amount to treating idioms as a special kind of word, which just happens to have spaces in it. But this is not a workable solution in general. A more reasonable idea is not to regard lexical lookup as a single process that occurs just once, before any syntactic or semantic processing, but to allow analysis rules to replace pieces of structure by information which is held in the lexicon at different stages of processing, just as they are allowed to change structures in other ways.

The second approach to idioms is to treat them with special rules that change the idiomatic source structure into an appropriate target structure. Clearly, this approach is only applicable in transfer or transformer systems, and even here, it is not very different from the first approach – in the case where an idiom is translated as a single word, it is simply a question of where one carries out the replacement of a structure by a single lexical item, and whether the item in question is an abstract source language word or a normal target language word.

One more problem with sentences which contain idioms is that they are typically ambiguous, in the sense that either a literal or idiomatic interpretation is generally possible. However, the possibility of having a variety of interpretations does not really distinguish them from other sorts of expression. Another problem is that they need special rules, in addition to the normal rules for ordinary words and constructions. However, in this they are no different from ordinary words, for which one also needs special rules. The real problem with idioms is that they are not generally fixed in their form, and that the variation of forms is not limited to variations in inflection (as it is with ordinary words). Thus, there is a serious problem in recognising idioms.

But the most significant problem of Russian-Byelorissian MT is homonymy.

The homonyms in these languages are formed as a result of [4]:

1. loss of semantic connection between separate values of the same word: *Месяц* – Earth satellite, *месяц* – 1/12 of a year; *свет* – universe, *свет* – aristocracy;
2. word transition from one part of speech into another (*столовая посуда* – adjective, *столовая открыта* – noun; *рабочы дзень* – adjective, *рабочы працуе* – noun; *вечером* – noun, *вечером* – adverb);
3. sound coincidence of the separate grammar forms for words with different meanings: *пила* – noun, *пила* – verb (*острая пила, пила чай*); *горка* – noun, *горка* – adverb (*невысокая горка, горка плакаць*);
4. sound coincidence of words borrowed from different languages: *кок* – a quiff, view of a hairdress (from French), *кок* – a ship's cook (from Dutch), *кок* (Russian *кокк*) – a bacteria in the form of small ball (from Greek).

Depending on the value and the way of formation homonyms are divided into two basic types:

- lexical, or simple homonyms;
- morphological, or derivative homonyms.

Russian and Byelorussian, as any Slavic language, are highly inflectional and almost free word-order languages. For example, most nouns or personal pronouns can form singular and plural forms in 7 cases. Most adjectives can form 4 genders, both numbers, 7 cases and 3 degrees of comparison.

The homonyms of the first type belongs to one part of speech and have identical inflections. Such homonyms always are marked in paper dictionaries by digits 1, 2, 3, etc., as different words.

The homonyms of the second type coincide on a sound structure only in the definite grammar forms. If they belong to different parts of speech such homonyms are called **homographs**. If they are different inflections for a word (noun, adjective, participle, etc) such homonyms are called **homoforms.**

If to consider the homonymy from the point of view of text production and perception it is possible to note the following:

- these is no homonymy for those who speaks or writes, if the writer specially does not want to use homonyms for any special purposes (doublemeaning, joke, etc.);
- there is almost no homonymy for those who hears and reads because a stress and a context remove practically all homonyms;
- but for MT the homonymy increases.

All these features create problems in translation for kindred languages.

2.4. Empirical Approach to MT

A huge number of techniques and computational approaches have been experimented in order to translate natural languages automatically. One of them is so-called *empirical* or *corpus-based* approach.

The increasing availability of large amounts of machine readable textual material has been seen by a number of research groups as opening possibilities for rather different MT architectures which apply relatively "low-level" statistical or pattern matching techniques either directly to texts, or to texts that have been subject to only rather superficial analysis. The reasoning behind the term empirical is that in such approaches, whatever linguistic knowledge the system uses is derived empirically, by examination of real texts, rather than being reasoned out by linguists [2].

Corpus-based approaches to MT have been on the rise recently, partly because of their promise to automate a great deal of dictionary construction and rule writing, partly because they simply represent a new way of attacking a stubborn problem, and partly because they have performed relatively well in MT evaluations. These approaches generally rely on a large bilingual text corpus to provide sample translations. A statistical model is trained on the samples, and it is used to translate new sentences. Corpus-based MT approaches have so far been applied to situations where large amounts of bilingual text already exist [5].

Recently, statistical data analysis has been used to gather MT knowledge automatically from parallel bilingual text: documents which are translations of each other. These techniques are extremely promising, as they provide a methodology for addressing the knowledge-acquisition bottleneck that plagues all large-scale NLP applications.

We offer to use an approach which is an attempt to avoid the drawbacks of traditional rule-based approaches and purely statistical approaches. Rule-based approaches, with rules induced by human experts, suffer from serious difficulties in knowledge acquisition in terms of cost and consistency. Therefore, it is very difficult for such systems to be scaled-up. Statistical methods, with the capability of automatically acquiring knowledge from corpora, are becoming more and more popular, in part, to amend the shortcomings of rule-based approaches. However, most simple statistical models, which adopt almost nothing from existing linguistic knowledge, often result in a large parameter space and, thus, require an unaffordably large training corpus for even well-justified linguistic phenomena.

In general, two kinds of features, namely statistical features and linguistic features (such as parts of speech and word senses) have been commonly used in various research works. Statistical features, such as mutual information and entropy, usually carry only statistical senses and carry few traditional linguistic notions. Linguistic features, such as parts of speech, on the other hand, are usually used to designate certain properties of the linguistic constructs under consideration.

Our methodology is based on the idea of constructing a corpus of virtual texts (CVT) [3] for both Russian and Byelorussian in order to receive parallel corpora for these languages. CVT is created from a corpus of initial texts T_0 with definite structural units of the given NL. A structural level of the language may be a level of a word, phrase, sentence, discourse, text. The finite set of texts of given NL has been chosen in accordance with particular criteria. From T_0 the source dictionary D_0 of word usages can be gathered.

Based on structural levels of NL we have developed classifier $K^{(r)}$ representing the whole system of grammatical, semantic, stylistic, phonetic, contextual and other parameters of Russian and Byelorussian. Both languages have the same classifier due to structural and lexical similarity.

A corpus of virtual texts together with program tools of access, extraction, analysis, etc. of NL information is a complex computer-assisted system providing solution to a wide range of information problems dealing with the study and use of the natural language. From the CVT we can receive various statictical characteristics for all structural levels of text and extract all possible linguistic rules existing in real text material.

Our purpose was to build Byelorussian/Russian bilingual text corpora in accordance with the suggested view on CVT. We have chosen the texts belonging to different subject areas and translated them in order to receive parallel texts. Each corpus has been morphologically analyzed, tagged,

lemmatized, and parsed. Semantic and grammatical features are indicated in accordance with predetermined classifier of such features $K^{(r)}$.

Based on the corpora we gathered bilingual dictionary, each part of which has the same structure due to the similarity of languages and may be represented as

$$D_0^{(r)} = \left\{x_i^{(r)}, K_i^{(r)}\right\}, i = \overline{1,m} \qquad (5)$$

where:

- $x_i^{(r)} \in D_0$, – a structural unit of the given NL of level r specified for each text;
- $K_i^{(r)} \subset K^{(r)}$ – a set of tags for the structural unit.

We tried to pursue the following objects:

1. Build a statistical toolkit and make it available to study. This toolkit includes corpus preparation software, bilingual-text training software, and run-time decoding software for performing actual translation.
2. Perform baseline evaluations. These evaluations consists of both objective measures (statistical model perplexity) and subjective measures (human judgments of quality), as well as attempts to correlate the two. We also produce learning curves that show how system performance changes when we vary the amount of bilingual training text.

We largely achieved these goals. We also had time to perform some unanticipated beyond-the-baseline experiments: speeding up bilingual-text training, using online dictionaries, and using language cognates. Finally, we built additional unanticipated tools to support these goals, including a sophisticated graphical interface for browsing word-by-word alignments, several corpus preparation and analysis tools, and a human-judgment evaluation interface.

We also used the toolkit as a platform for experimentation. Our experiments included working with distant language pairs (such as Byelorussian/Russian), rapidly porting to new language pairs, managing with bilingual data sets, speeding up algorithms for decoding and bilingual and text training, and incorporating morphology, syntax, dictionaries, and cognates.

We can consider our approach as compromise between the two extremes of the spectrum for knowledge acquisition. This approach emphasizes use of well-justified linguistic knowledge in developing the underlying language model and application of statistical optimization techniques on top of high level constructs, such as annotated syntax trees, rather than on surface strings, so that only a training corpus of reasonable size is needed for training and long distance dependency between constituents could be handled.

3. Conclusion

We have described a methodology which demonstrates how to solve the problems aring during Russian-Byelorussian MT.

The results reached for the corpus are very good and promising. Larger amount of data can significantly help the MT. As the general translation tool has been just developed, it is now possible to experiment with different system parameters, such as the number of iterations of particular models, and to adjust the translation models to better suit the Russian/Byelorussian language pair.

References

[1]. Douglas Arnold, Lorna Balk, Siety Meijer, R.Lee Humphreys, Louisa Sadler. *Machine Translation: an Introductory Guide*. – NCC Blackwell, London, 1994. – 238 p.

[2]. Roubashko N.K. Development and Representation of Linguistic Knowledge for Natural Language Processing, *International Conference on Systems and Signals in Intelligent Technologies* (SSIT'98). Minsk, 1998. – P.383–389.

[3]. Roubashko N.K. Word-Sense Disambiguation: a Corpus-Based Approach, *Sixth International Conference "Pattern Recognition and Information Processing"* (PRIP'2001). – Minsk, 2001. – Vol.2. – P. 187–191.

[4]. Разработать программные средства машинного перевода текста (деловая проза): Отчет о НИР (заключ.) БГУ; рук. работ Совпель И.В.; № ГР 1994589. – Минск, 1997. – 79 с.

[5]. Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, David Yarowsky. *Statistical Machine Translation*: Final Report. – JHU Workshop, 1999. – 42 p. – Available at http:// citeseer.nj.nec.com/al-onaizan99statistical.html