

# **An Experimental Latvian-English Direct Machine Translation System**

Normunds Grūzītis, Gunta Nešpore, Baiba Saulīte

Nordic Graduate School of Language Technologies  
Machine Translation

January 8, 2007

## **1. Introduction**

The report describes a simplistic machine translation system, which has been developed to fulfill requirements of the Assignment #1 of the Machine Translation course at the Nordic Graduate School of Language Technologies.

Basic requirements for the implementation: sentence splitting, tokenization, handling of capital letters, dictionary look-up and lexical substitution, heuristics for handling ambiguities, copying unknown words, digits, signs of punctuation etc.

## **2. Latvian in a Nutshell**

Latvian belongs to the Baltic language group — it is highly inflective synthetic language with rather free word order. Due to the fact that there is virtually no language with completely free word order and vice versa we are claiming that Latvian has one of the most liberal word ordering. Its grammar structure it is closely related to Lithuanian and also to Slavonic languages (int. al. many Central European languages).

In terms of morphology Latvian is mainly suffixing language. Nouns inflect for number and case. A nominal word-form usually consists of a stem plus inflectional ending. Endings may be polysemous (e.g. gender + number + case) and homonymous (e.g., the same ending stands for both SG GEN and PL NOM). Due to inflectional classes (declensions), there are also synonymous endings (e.g. masculine NOM endings -s, -š, -is, -us). Each noun has fixed grammatical gender, either masculine or feminine.

### 3. System Architecture

As it follows from the title “direct MT”, the model is based on word-by-word processing and translation:

- 1) Input sentence is split into a list of tokens.
- 2) For each token:
  - a) find lemma (base form) via morphological analysis;
  - b) look for possible substitutions of the source lemma in a bilingual lexicon;
  - c) synthesize an appropriate form of the target lemma, according to the morphological features of the corresponding token;
  - d) apply additional grammar rules on the target word form, if appropriate;
  - e) attach the resulting set of translation equivalents to the source token.
- 3) Construction (concatenation) of a target sentence; if at least one token has more than one translation equivalent attached, there is possible more than one reading of the target sentence.
- 4) Disambiguation, if necessary.

The framework itself is rather robust and could be applied to certain language pairs other than Latvian-English as well.

#### 3.1. Lexicons

There are three lexicons used in our approach:

- 1) Lexicon that enumerates possible word forms of the source language (Latvian) along with their morphological features and base forms:

Word form	Features	Lemma
vilciena	n, m, sg, gen	vilciens
vilcienā	n, m, sg, loc	vilciens
vajadzēja	v, past	vajadzēt
nofotografēt	v, inf	fotografēt
dziļajā	adj, m, sg, loc, def	dziļš
dziļā	adj, m, sg, loc, indef	dziļš

- 2) Lexicon that enumerates pairs of translation equivalents (in terms of base forms):

Lemma — source language	Lemma — target language
vilciens	train
vajadzēt	have
vajadzēt	need
fotografēt	take pictures

- 3) Lexicon that enumerates possible word forms of the target language (English) along with their morphological features and base forms:

Word form	Features	Lemma
train	n, 0, sg, 0	train
had	v, past	have
will have	v, future	have
to take pictures	v, inf	take pictures

For the sake of simplicity, we have declared only those features that are exploited in agreement between source and target word forms or used by grammar rules, as well as few others as illustration.

The idea is neither to hard-code the morphotactics and orthographic rules nor to use a morphological analyzer. Instead we prefer to enumerate and annotate all the forms explicitly. Consequently morphological analysis/synthesis is reduced as a search problem, which is simple and efficient method. Moreover, it is easier to implement a reliable synthesizer (in order to acquire the lexicons) than an analyzer (especially for highly inflective languages).

There are several shortcomings, which we have solved only partially, namely, collocations and mappings of morphological features over the languages.

It is a common situation that there doesn't exist 1:1 translation equivalents — the same concept in one or the other language can be represented by a collocation (1:n or n:n translations, e.g. *fotografēt* = *take pictures*). Or even more — the same concept in both languages can be represented by collocations (n:n). If we are not considering the practical aspects, it is not a problem to put such mappings into the tables. However, in our framework it is allowed to use collocations only in the target language, as it doesn't impact the simplistic word-by-word processing.

Another problem is caused by translation equivalents, which have different morphological features in each language. For example, let's consider following lemmas: *māja* (sg) = *house* (sg), *mājas* (pl) = *houses* (pl), however, the homonym *mājas* (pl-only) = *home* (sg-only). Other

examples that are present in our test corpus are: *svētdien* (adv) = [on (prep)] *Sunday* (n) and *kas* (pron) = *that* (conj).

Although such words can't be fully mapped between languages (in terms of their features), it seems that there are no problems at the level of bilingual dictionary, so at least the right target lemma will be acquired.

### 3.2. Grammar Rules

The very basic rule already discussed in the previous section is that source and target word forms have to agree on their morphological features, i.e. the target word form has to conform to the source word form. If a part-of-speech in one language has a feature, that doesn't exist for the part-of-speech in the other language then a null-value have to be put in the particular position (like {n,o,sg,o} and {adj,o,o,o,o} in English).

Inflective (synthetic) languages on the one hand encode a lot of information in word forms. Analytic languages on the other hand have to encode the same information via analytic forms (e.g. prepositional phrases). We have defined few simplified rules that demonstrate how it would be possible to deal with some of the mappings:

1.  $wf_L \{n, \_, \_, loc\} \rightarrow in \{prep\} + wf_E \{n, o, \_, o\}^1$   
 $wf_L \{adj, \_, \_, loc, \_\} \rightarrow in \{prep\} + wf_E \{adj, o, o, o, o\}$   
 $wf_L \{v, participle, loc, \_\} \rightarrow in \{prep\} + wf_E \{v, participle, o, o\}$   
 e.g. *pasakā* {n,f,sg,loc}  $\rightarrow in + fairytale$ 
  - If the previous token matched one of the patterns and so do the current token, then the substitution at the current position is not applied.  
 e.g. *skaistā* {adj,f,sg,loc,indef} *pasakā* {n,f,sg,loc}  $\rightarrow in + beautiful\ fairytale$
2.  $wf-1_L \{n, \_, \_, gen\} wf-2_L \{n, \_, \_, \_\} \rightarrow wf-2_E \{n, o, \_, o\} + of \{prep\} + wf-1_E \{n, o, \_, o\}$   
 e.g. *saules* {n,f,sg,gen} *staros* {n,m,pl,loc}  $\rightarrow in + rays + of + sun$
3.  $wf_L \{adj, \_, \_, \_, def\} \rightarrow the \{art, def\} + wf_E \{adj, o, o, o, o\}^2$   
 $wf_L \{v, participle, \_, \_, def\} \rightarrow the \{art, def\} + wf_E \{v, participle, o, o\}$   
 e.g. *klusais* {adj,m,sg,nom,def}  $\rightarrow the + calm$ 
  - If the previous token matched one of the patterns and so do the current token, then the substitution at the current

<sup>1</sup> Where *wf* stands for 'word form', L for 'Latvian', E for 'English' and '\_' for any feature.

<sup>2</sup> Where *def* stands for 'definite'.

position is not applied.

e.g. *klusajā* {adj,m,sg,loc,def}, *skaistajā* {adj,m,sg,loc,def}

*mežā* {n,m,sg,loc} → *in + the + calm, beautiful forest*

We haven't implemented the substitution of a genitive phrase to a preposition-*of* phrase in our system, as it requires to alter the word order, as well as sg-gen in Latvian is morphologically ambiguous with pl-nom.

Appropriate usage of indefinite (*a*) and definite (*the*) articles in general could be possible if discourse analysis and anaphora resolution in particular would be performed. In a shallow manner, the only case of which we can be quite sure about is usage of definite articles before definite noun phrases (definite adjective or participle + noun).

Another class of grammar rules should consider the usage of commas. In Latvian there are strict rules when a comma must be used, in English it seems not to be so strictly defined. Nevertheless, syntactic parsing should be applied in order to decide whether an existing comma should be retained and whether a new comma should be introduced.

### 3.3. Disambiguation

There are two basic types of ambiguities: grammatical (morphological and syntactical) and lexical. Due to the fact that our system do not perform syntactic parsing it is impossible to determine, which parts of a sentence are dependants of the other parts thus we can't establish agreement on morphological features (e.g. number, case) between parts of a sentence. The consequence is morphological ambiguity and the system has to propose more than one variation of the translation. Disambiguation is left for the reader.

*Meitene izgāja no mājas.*

*Girl went out form house.*

← *mājas* {n,f,**sg,gen**}

*Girl went out form houses.*

← *mājas* {n,f,**pl,nom**}

In some cases trivial heuristics may be applied: if two sets of features differ only in terms of a gender, it can be ignored as there will be no difference in English. In general, if it turns out that there are equal strings (sentences) in the resulting set of translations, we can remove the redundant ones.

Concerning the lexical ambiguities, we do not apply any shallow word sense disambiguation by default. Of course, we could apply most

frequent sense heuristics or some unsupervised method like Lesk's algorithm (based on definitions acquired from a monolingual dictionary of Latvian), but this wouldn't give much improvements. Moreover, we believe that premature disambiguation has more shortcomings than benefits. However, we have implemented an optional feature, which simulates the most frequent sense heuristics and can be activated by a user — random selection of each ambiguous translation equivalent, as there is no particular ordering in the lexicon.

It should be noted that in case of manual disambiguation, translation process should be limited to a single sentence at a time, i.e. sentence by sentence processing asking for the most appropriate reading in case of ambiguity. Otherwise the user will be overloaded with all the possible variations. For instance, first four sentences of our small test corpus using the current, very limited lexicon produces 24 translation variations, but translation of the full text at a time causes *java.lang.OutOfMemoryError* — 1 572 864 possible variations!

Meitene izgāja no mājas. Bija skaista un saulaina ziemas diena. Viņai noteikti vajadzēja noķert nākamo vilcienu, kas viņu aizvedīs uz piesnigušo mežu. Kamēr vēl gaišs, viņa gribēja nofotografēt baltās egles un priedes, un kluso, skaisto mežu. Vilcienā meitene apsēdās un lasīja dienas avīzi, kur bija rakstīts, ka ziema ir sākusies pēkšņi. Vilciens brauca caur netīro, dubļaino pilsētu un ievada pasakā. Mežs bija brīnumains. Tur, kur viņa vasarā lasīja sēnes un ogas, tagad gulēja sniega kalni. Egļu zari liecās zemē, jo sniegs bija smags. Viss bija balts un mirdzēja saules staros. Meitene iebrida dziļajā sniegā un fotografēja. Tur bija saules stari sniegā, dzenis kokā, piesnigušie bērzu zari. Debesis bija daudzkrāsainas un saule bija ļoti skaista. Pēc mirkļa saule pazuda un sāka snigt. Meitene paslēpa fotoaparātu somā un skrēja atpakaļ uz vilciena pieturu. Ceļā uz mājām visu laiku snīga. Meitene domāja par skaisto dienu. Pasaule šķita tīra un balta kā sapnī.

### 3.4. Pre & Post Processing

Preprocessing involves tokenization of an input sentence. If a text (more than one sentence) has been submitted, sentence splitting comes for free via tokenization, as we are not dealing with syntax, so we don't care about borders of a sentence.

Postprocessing deals with (i) combination and concatenation of translation equivalents together with prepositions and articles that have been introduced (if any), (ii) retaining of capital letters and (iii) highlighting of words that were not recognized (not present in at least one of the tables: morph-lv, lemmas, or morph-en) or were not fully mapped between both languages.

### 3.5. User Interface

The application is available at <http://www.ailab.lv/Normunds/MT>. It is possible to submit a sentence or a short text (carriage return/new line characters are not supported). All the possible target readings are printed out. If the checkbox “I’m lucky” is selected, only one translation equivalent per token will be selected (in a random manner).

For experimental purposes it is possible to add new entries in all the tables of the lexicon on-line, as well as browse the current content of the lexicon.

## 4. Evaluation

In general, there does not exist only one correct translation. One variation is the translation that is done by the system – for complicated sentences the quality is not at all the best. Another variation is the “golden-standard” corpus. But there exist some more possibilities to translate the sentence. If the translation of the system does not correspond to the translation of the sample corpus, it does not necessarily mean that it is wrong.

System can translate simple sentences quite well where word order in Latvian is very close to English. Below is the best translation of the sample text achieved by our system, exploiting the current lexicon and grammar rules, and the guidelines, which were proposed in the section 3.3.

<p>Girl went out from house. Was beautiful and sunny winter day. She definitely had to catch the next train, that she will bring to the snow-covered forest. While still light, she wanted to take pictures the white fir-trees and pines, and the calm, beautiful forest. In train girl sat down and read day newspaper, where was written, that winter is begun suddenly. Train went through the dirty, muddy city and brought in fairytale. Forest was wonderful. There, where she in summer gathered mushrooms and berries, now lay snow hills. Fir-trees branches leaned down, because snow was heavy. Everything was white and glittered sun in beams. Girl waded in the deep snow and took pictures. There were sun beams in snow, woodpecker in tree, the snow-covered birches branches. Sky was colorful and sun was very beautiful. After moment sun disappeared and began to snow. Girl hid camera in bag and ran back to train stop. In way to home everything time snowed. In way to home all time snowed. Girl thought about the beautiful day. World seemed clean and white like in dream.</p>
---

In many cases in the English translations lacks articles, because often regular connections between Latvian morphological (Latvian does not have articles) features and English articles can be established. Though in

cases, when morphological connections are straightforward, translations have the correct definite articles.

Latvian case forms correspond to English prepositional forms. Therefore there are rules in the system saying that Latvian locative corresponds to construction with *-in* in English (*sniegā – in snow*). In some sentences these rules is not sufficient, and we get a translation *Sun in beam* (instead of *in beams of the sun*). To obtain the correct translation the rules should be extended and the possessive forms (it changes word order) should be described.

In the sentences with different word order in Latvian and English the translation is not very well, because the syntactic analysis is not done and the rules for changes of word order are not made. There are some difficulties with pronouns that are not used in the same case. Both of these problems can be seen in the fragment that is translated as *that she will take to the snow-covered forest* (instead of *that will take her to the snow-covered forest*).

In some cases translation is not complete (it lacks the grammatical subject, like *it is*), because in Latvian such a subject is not mandatory; of course it does exist in the deep structure, but in the text it can be skipped, that is why we got translations like *while still light* (instead of *while it is still light*).

An informal criterion for evaluation of an MT system, which is nothing more than a direct system, could be as follows: if a user can comprehend the main idea out of an arbitrary text, then system performs reasonably.

## 5. Potential Improvements

Syntactic parsing would eliminate virtually all morphological ambiguities and it would be possible to generate smoother sentences (analytic forms, word ordering), however, syntactic parsing itself, in general, is highly ambiguous – the number of possible parse trees is expected to be higher than the number of morphological ambiguities (it depends on the granularity of the grammar). Besides, it would be necessary to introduce additional syntax-mapping rules from one language to another.

State-of-the art achievements in shallow word sense disambiguation show that it might be possible to reduce ambiguity only for about 70% for an arbitrary text (base line + 30%). Stochastic methods heavily depend on the size and structure of the learning corpus. They turn out to be more or less “magical” calculations, a black-box, which doesn’t care a

lot about the knowledge of a language and the sense that is expressed via language structures. We would prefer to choose a hybrid approach instead: when rule-based methods are done, the remaining syntactic ambiguities (all the possible head-dependant pairs) could be reduced exploiting statistics of selectional restrictions extracted from a corpus.

## **6. Conclusions**

Sentences of a simple structure, limited usage of analytic forms and word ordering similar to English are translated rather acceptably for such a simple system, however, for real, arbitrary texts (like the bonus sentence in the test corpus) this approach is a deadlock.

If we would start to implement the potential improvements discussed in the previous section, then this will not be a [simplistic] direct MT anymore.