A Study of Indonesian-to-Malaysian MT System

Septina Dian Larasati, Vladislav Kuboň
Inst. Of Formal and Applied Linguistics
Charles University
Prague, Czech Republic

Abstract—**The paper presents an ongoing work on the implementation of an MT system between Indonesian and Malaysian. The system uses a method of almost a direct translation exploiting the similarity of both languages. This method was previously used on a number of language pairs of European languages. The paper also makes an overview of linguistic phenomena which can negatively influence the translation quality and it suggests a solution for some of them.**

Keywords-machine translation; related languages; direct translation; morphology; hybrid method

I. Introduction

Probably none other linguistic application area has attracted as much research effort as the area of automatic translation of texts between natural languages (a field usually called Machine Translation -- MT). After more than fifty years of research during which there were periods of uncritical expectations followed by long periods of bitter despair, the application of stochastic methods brought new hopes into a field which notoriously failed to provide acceptable results. The stochastic methods rejected traditional rule-based approaches and replaced them by the exploitation of bigger and bigger amounts of data. The lack of large coverage grammars was replaced by a lack of parallel data.

Although nowadays the expectations are yet again very high, it is clear that not even the current breakthrough caused by stochastic or hybrid approaches as, e.g., in the factored translation model described in [17], will solve all the problems, especially the problems of less represented languages.

One property which makes the translation task easier is the relatedness of the source and target languages. The relatedness usually means a great deal of similarity at all levels, but the experiments carried out in the past (cf. the references further in the text) have shown that the most important level is the level of syntax closely followed by morphology.

This article describes an experiment with the application of an existing model for the MT between related languages on a new language pair from a very different language group. The architecture of the system is based primarily on rule-based approach which allows for a great deal of ambiguity in all steps. This ambiguity is then resolved by a simple stochastic ranking of all translation hypotheses. The simple architecture was originally developed for European languages and one of the main goals of this paper is to describe the issues encountered in the process of the application of the method to a pair of Asian languages which are typologically different from the European languages for which the method has been originally developed (Slavic and Romance languages).

If we look at the experiments made so far for related languages, we will find numerous experiments which have been performed recently for various language groups:

- for Slavic languages in [12] and [16],
- for Scandinavian languages in [3], [6], and [13],
- for Turkic languages in [10]
- and for languages of Spain in [1].

The close relatedness of natural languages from one typological group (and sometimes even across the group borders, cf., Czech-to-Lithuanian experiment described in [8]) makes the translation task easier thus allowing for the application of methods which would not be good enough for the translation of unrelated language pairs. Using simpler methods does not mean a lower translation quality - many of the translation errors result from the imperfect attempts to parse a source language fully, in some cases even to the deep syntactic level of representation. The accumulation of errors in parsing, transfer and generation in the systems using the classical transfer-based architecture substantially decreases the translation quality.

II. Typology of the Language

Although spoken by millions of speakers, research on this pair of languages has not been very enthusiastic compare to most of the European languages. This makes these two closely related languages under question very compelling to be explored. Coming from the same language family, Austronesian, the languages share similar behavior which usually being misapprehended by non-natives that they both are mutually intelligible. The languages are very dynamic where the evolution makes them differ from one another.

Both of these agglutinative languages have similar morphology mechanisms and share some words, both the words with exact or similar meaning and also the words with

This work was supported by an Erasmus Mundus Master Program Language and Communication Technologies and partially supported by the grant MSM 0021620838 of the MŠMT ČR.

different meaning that can be misinterpreted by both native speakers. Example on words that can be misinterpreted is the word '*kereta*' which means 'car' in Malaysian and 'train' in Indonesian. That word can be inflected in the same way such as '*berkereta*' which means '*having car*' in Malaysian and '*having train*' in Indonesian. With these backgrounds, this language pair is a suitable pair to apply this shallow rule-based MT method.

Orthography – The alphabet is basic modern Latin alphabet with hyphen used to separate words on the reduplication case and on special clitic case.

Word Order – The word order is fixed and the position in the sentence is essential to determine the role of the word in the sentence.

Tense – The languages do not have special inflection tense marking. The tense are marked by using additional word or temporal information in the sentence.

Voice – The sentence voices are marked by different prefix of the inflected word.

Gender – Classification of gender is not common although it occurs in some irregular cases marked by several suffixes. This fashion is now rarely used and not productive any longer.

Number – The plurality is not only found in Nouns but also in other Part-of-Speech (POS) where it marks the plurality of the action or referring to plural entities.

III. Architecture of the System

Most of the systems mentioned in the introduction section try to exploit the similarity of closely related languages. This can apparently be done only in case that the system architecture is reasonably simple. The more complicated the architecture is, the higher number of errors is introduced into the translation process by individual modules. These errors then negate the advantage of working with closely related languages.

The most successful architecture for simple MT systems had been developed for the system Česílko [7], and also used by the system Apertium [1]. The fact that Apertium is an open-source platform and thus can easily been adopted for experiments with other language pairs led us to the decision to use it for our experiments with two South-Asian languages, Malaysian and Indonesian.

As mentioned above, the architecture of Česílko and Apertium is relatively simple. The systems are in fact transfer based systems with the transfer being performed either at the morphological or shallow syntactic level (depending on the degree of syntactic similarity of a source and a target language). The role of morphology in such a system is really crucial.

Indonesian and Malaysian MT system is implemented on Apertium (http://www.apertium.org), a free/open-source MT platform for developing rule-based machine translation system [15]. This platform is a shallow-transfer machine translation engine word-to-word machine translation to produce fast, reasonably intelligible and easily correctable translations not only between related languages but also can be extended for language pairs which are not closely related.

Apertium has a modular architecture [2] and in each module it provides various tool options depending on the nature of the language. In this MT system some module are skipped from the original setting. The modules that are being kept in this MT system are

Morphological Analyser – the surface forms are segmented and each form will be analyzed to get the lexical unit, such as lemma, Part-of-Speech tags and morphological inflection information. Apertium offers various morphological analysis tools that can accommodate different nature of languages. For this particular language pair under question, the morphological analyser are developed based on Xerox finite-state tools (XFST) and high-level declarative language to specify language lexicon (LEXC), which then compiled in Foma (http://foma.sourceforge.net/) [14], a finite state toolkit that implements Xerox xfst and lexc. This module includes the source language monolingual dictionary as well.

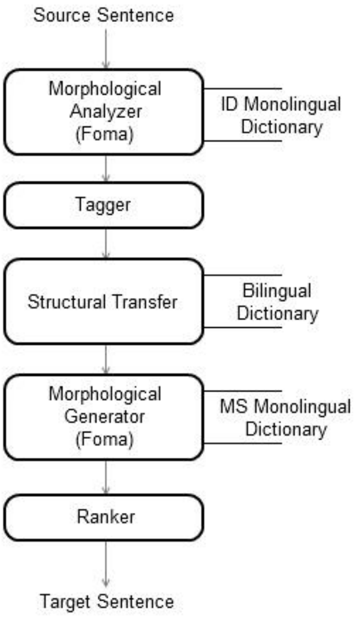
Source Sentence
Morphological Analyzer (Foma)
ID Monolingual Dictionary
Tagger
Structural Transfer
Bilingual Dictionary
Morphological Generator (Foma)
MS Monolingual Dictionary
Ranker
Target Sentence

Figure 1. MT System Modular Architecture

Part-of-Speech Tagger – trained using text corpus and tagger definition file to disambiguate the analysis.

Lexical Transfer and Structural Transfer – reads each source language word analysis and transfers it into the target language using bilingual dictionary. Structural transfer between source and target language can be done in three stages, Chunker, Interchunk, and Postchunk depending on the need. This MT system only utilizes one stage transfer.

Morphological Generator – the reverse direction of Morphological Analyser to generate the analysis results to their surface forms.

Ranker – is also added to choose the best translation hypotheses statistically.

IV. Morphological Analysis and Generation

Considering the typology of the languages under question, the extensive engineering task falls on the morphological analyser and generation compared to the other parts. Here describes the morphological operations of the language followed by how the analysis and generation are implemented.

A. Morphological Operations

The language pair has similar morphological mechanism. We broke down this mechanism into four morphological operations. Those operations that have to be handled are

1. *Affixation*. This operation including prefix, suffix, and circumfix. There are several cases of infixes, which now are rarely used. These special cases are being handled differently in the language resource part (see *Language Resource*).
2. *Reduplication*. The reduplication can occur on any POS. It is divided into three different types, namely full reduplication, partial reduplication and affixed reduplication. Partial reduplication is not handled in the morphological analyser but treated as an entry in the dictionary.
3. *Clitic*. Enclitic and proclitic are representing the pronouns. It can be kept as clitic or restored to its corresponding independent pronoun, where both ways are grammatically correct.
4. *Particle*. Particle marks the stress, level of formality and constructing question words.

Shown in Figure 2, the schema of how the inflection around the lemma. The prefix itself is divided into two depending on the position and then named as pre-prefix and prefix. The reduplication can occur almost everywhere in the affixed lemma.

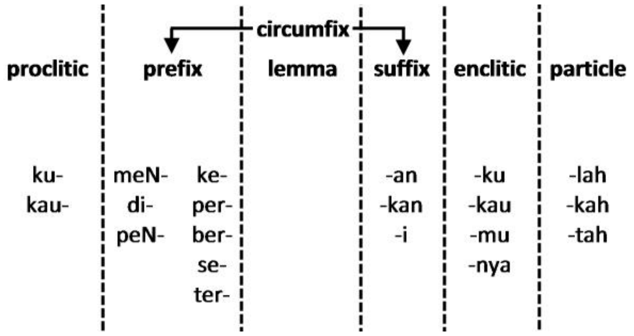

Figure 2. Morphological Operations Schema

B. Morphological Tool

Since the morphological mechanisms are similar, we simply use the same morphological analyser for both languages. The widely used tool to do analysis and generation on Apertium platform is Lttoolbox, a toolbox for lexical processing, morphological analysis and generation of words. This tool has been used on several language pairs and mostly on languages that has the inflection on suffix as Apertium was initially designed for. It works by defining exhaustive combination of the inflection forms that are possible in a language, called paradigm. We found that this tool cannot accommodate well Indonesian and Malaysian morphology by these several limitations:

- *The treatment for morphemes that precedes the base word is not straightforward.* The analysis expected from this module is in the form of lemma followed by morphological tag(s), for example `pesan<n><bare><sg>`. The process of the analysis is done on the position of the inflection. Therefore the prefix analysis, which is the tag(s), will be in the front of the lemma. By this, a separate additional reformatting needs to be done. Moreover, circumfix will be treated as independent prefix and suffix.
- *The morphophonemic are handled by expanding the morpheme to its whole possible inflection forms.* For example for the pre-prefix 'meN-' will be expanded to its several different forms considering to which base word it glued to. This morpheme will inflect into 'menge-' for one syllable case, 'meng-' for words starting with [a i u e o g h], 'meny-' for words starting with [s, y] and so on.
- *This tool cannot handle reduplication cases.*

Therefore to encounter this we decided not to use Lttoolbox and initially employed an available Indonesian morphological analyser [4], which was developed in xfst/lexc platform. This tool has already handled the reduplication and Indonesian morphophonemic. To incorporate this tool to Apertium we compiled it in Foma, a finite-state toolkit.

This morphological analyser includes large number of Indonesian lemmata, but unfortunately the coverage of how it handles the inflections was not adequate enough for the task, where

- *It covers partly the morphological operations.* The morphological operation that it handles was reduplication and several affixations, not including clitic and particle. The uncovered cases will cause the inflected word to be left un-translated.
- *The tagset is underspecified for generation.* It consists of 17 general tags, which mostly tag the Part-Of-Speech (POS) and the morphological operation that occurs. The POS tag simply marks three POS types, namely Verb, Noun, and Adjective, while others are considered as Etc.
- *Several inflected words have the same analysis*, which is unfavorable for the translation since those different inflected words will be transferred to the same target analysis. For example in the case of the noun derivation '*kiriman'*, '*pengirim*' and '*pengiriman'* from the verb '*kirim*' will have `kirim+Noun` as the result of the analysis.

- Yet relating to the tagset problem, *the generation step generates a big number of inflected words*, which will produce bigger numbers of translation hypotheses. For example, the analysis `kirim+Noun` will generate words as showed in Table I.

TABLE I. PROBLEM IN THE ANALYSIS/GENERATION

Analysis Result		
kiriman pengirim pengiriman	>	kirim+Noun
Generation Result		
kirim+Noun	>	pengirim pengiriman *pemberkiriman *perkiriman *kepengiriman *keberkiriman *kekiriman kiriman

*) marks the ungrammatical inflected words

#) marks the un-generated inflected words

Initiating from that we take the part where it handles the morphophonemic and reduplication. Then we build a morphological analyser with more extensive inflection coverage. We also introduce more fine-grained tags and change the forms from +TAG into <TAG> to suit Apertium platform.

TABLE II. MORPHOLOGICAL TAGSET

Tag	Description	Tag Type
<adj>	adjective lemma	POS
<n>	noun lemma	POS
<num>	number lemma	POS
<prn>	pronoun	POS
<det>	determiner	POS
<cnjcoo>	coordinating conjunction	POS
<cnjsub>	subordinating conjunction	POS
<vblex>	verb lemma	POS
<part>	particle	POS
<mod>	modal	POS
<ij>	interjection	POS
<qst>	question word	POS
<pr>	preposition lemma	POS
<p1>	first person	PERSON
<p2>	second person	PERSON
<p3>	third person	PERSON
<sg>	singular	NUM
<pl>	plural	NUM
<card>	cardinal number	DERNUM
<ord>	ordinal number	DERNUM
<coll>	collective number	DERNUM
<ref>	referential number	DERNUM
<vbhaver>	verb ‘to have’	VERBVAR
<vbser>	verb ‘to be’	VERBVAR
<actv>	active voice	VOICE
<pasv>	passive voice	VOICE
<perf>	perfective aspect	ASPECT
<imp>	imperfective aspect	ASPECT
<bare>	bare noun	DERNOUN
<abstract>	derived abstract noun	DERNOUN
<actio>	derived action noun	DERNOUN
<actor>	derived actor noun	DERNOUN
<ent>	derived entity noun	DERNOUN
<theme>	derived theme noun	DERNOUN
<positive>	bare adjective	DERADJ
<sup>	superlative adjective	DERADJ
<exceed>	adjective that shows something exceeding	DERADJ
<manner>	adjective that shows similar manner	DERADJ
<uni>	union adjective	DERADJ
<possib>	adjectival phrase	DERADJ
<enc>	enclitic	CLITIC
<pro>	proclitic	CLITIC
<appl>	applicative	TRANSITIVITY
<caus>	causative	TRANSITIVITY
<cap>	capitalization mark	MARK
<pos>	possesive mark	MARK

Comparing to the previous example, with the current morphological analyser the analysis are more precise.

TABLE III. CURRENT ANALYSIS/GENERATION

Analysis Result		
kiriman	>	kirim<vblex><ent><sg>
pengirim	>	kirim<vblex><actor><sg>
pengiriman	>	kirim<vblex><actio><sg>
Generation Result		
kirim<vblex><actor><sg>	>	pengirim
kirim<vblex><actio><sg>	>	pengiriman *#pemberkiriman *#perkiriman *#kepengiriman *#keberkiriman *#kekiriman
kirim<vblex><ent><sg>	>	kiriman

*) marks the ungrammatical inflected words

#) marks the un-generated inflected words

Here is the analysis for Indonesian sentence “*apabila, sebelum mengunduh, menginstal, mengaktifkan atau menggunakan piranti lunak, anda memutuskan bahwa anda tidak bersedia untuk menyetujui ketentuan-ketentuan perjanjian ini, anda tidak bisa dan tidak berhak menggunakan piranti lunak ini*” (“if, before downloading, installing, activating or using the software, you decided that you are unwilling to agree to this agreement terms, you cannot and do not have right to use this software”).

```
^apabila/apabila<cnjsub>$
,
^sebelum/sebelum<cnjsub>$
^mengunduh/unduh<vblex><actv><imp><sg>$
,
^menginstal/instal<vblex><actv><imp><sg>$
,
^mengaktifkan/aktif<adj><actv><imp><caus><sg>$
^atau/atau<cnjcoo>$
^menggunakan/guna<n><actv><imp><caus><sg>$
^piranti~lunak/piranti~lunak<n><bare><sg>$
,
^anda/anda<prn><p2><sg>$
^memutuskan/putus<adj><actv><imp><caus><sg>$
```

```
^bahwa/bahwa<cnjsub>$
^anda/anda<prn><p2><sg>$
^tidak~bersedia/enggan<adj><positive>$
^untuk/untuk<pr>$
^menyetujui/setuju<vblex><actv><imp><appl><sg>$
^ketentuan-ketentuan/tentu<adj><abstract><pl>$
^perjanjian/janji<n><theme><sg>$
^ini/ini<det>$
,
^anda/anda<prn><p2><sg>$
^tidak/tidak<adv>$
^bisa/bisa<mod>/bisa<n><bare><sg>$
^dan/dan<cnjcoo>$
^tidak/tidak<adv>$
^berhak/hak<n><actv><perf><vbhaver><bare><sg>$
^menggunakan/guna<n><actv><imp><caus><sg>$
^piranti~lunak/piranti~lunak<n><bare><sg>$
^ini/ini<det>$
```

Figure 3. Analysis Example for Indonesian Sentence "*apabila, sebelum mengunduh, menginstal, mengaktifkan atau menggunakan piranti lunak, anda memutuskan bahwa anda tidak bersedia untuk menyetujui ketentuan-ketentuan perjanjian ini, anda tidak bisa dan tidak berhak menggunakan piranti lunak ini*"

The generation process is simply the opposite direction of the analysis, where the surface forms are composed based on the analysis.

V. Disambiguation

Although the morphological analysis has been expanded to prevent ambiguities, but cases such as homophones will still remain. The word *'bisa'* in the previous analysis example (Figure 3) will have two possible analyses since it is a homophone for the word 'can/able to', a modal verb, and 'snake venom', a noun. This several analyses are disambiguated statistically based on some probability.

The disambiguation of the analyses is done in the POS tagger. There are several ways provided by Apertium to train the Tagger. We choose to use the target language tagger training, that provided by Apertium [5]. This training process is relatively faster and more suitable for our MT system which only has one-stage transfer. It trains the tagger based on the source and target language. Intend to do that we need to have a text corpus in source and target languages, a tag definition file, and having the MT system running. In the tag definition file we specify the sequence of tags that is enforced or forbidden to be occurring in the analysis. The analysis of the word *'bisa'* in Figure 3 is being disambiguate into

```
^bisa<mod>$
```

VI. Transfer

The translation to the target language takes place in the lexical and structural transfers. The analyses of the source language are transferred into the target language and then it is generated to the target surface form.

The transfer between the two languages is done using transfer rules and bilingual dictionary. The sentence structure of both languages is similar where reordering is not required. We use Lttoolbox to keep the bilingual dictionary.

```
<e><p><l>apabila<s n="cnjsub"/></l>
       <r>jika<s n="cnjsub"/></r></p></e>
<e><p><l>sebelum<s n="cnjsub"/></l>
       <r>sebelum
              <s n="cnjsub"/></r></p></e>
<e><p><l>unduh<s n="vblex"/></l>
       <r>muatturunkan
              <s n="vblex"/></r></p></e>
<e><p><l>instal<s n="vblex"/></l>
       <r>pasang<s n="vblex"/></r></p></e>
<e><p><l>aktif<s n="adj"/></l>
       <r>aktif<s n="adj"/></r></p></e>
<e><p><l>atau<s n="cnjcoo"/></l>
       <r>atau<s n="cnjcoo"/></r></p></e>
<e><p><l>guna<s n="n"/></l>
       <r>guna<s n="n"/></r></p></e>
<e><p><l>piranti~lunak<s n="n"/></l>
       <r>perisian<s n="n"/></r></p></e>
<e><p><l>anda<s n="prn"/></l>
       <r>anda<s n="prn"/></r></p></e>
<e><p><l>putus<s n="adj"/></l>
       <r>putus<s n="adj"/></r></p></e>
<e><p><l>bahwa<s n="cnjsub"/></l>
       <r>bahawa<s n="cnjsub"/></r></p></e>
<e><p><l>enggan<s n="adj"/></l>
       <r>enggan<s n="adj"/></r></p></e>
<e><p><l>untuk<s n="pr"/></l>
       <r>untuk<s n="pr"/></r></p></e>
<e><p><l>setuju<s n="vblex"/></l>
       <r>bersetuju
              <s n="vblex"/></r></p></e>
<e><p><l>tentu<s n="adj"/>
              <s n="abstract"/>
              <s n="pl"/></l>
       <r>terma<s n="n"/><s n="bare"/>
              <s n="pl"/></r></p></e>
<e><p><l>janji<s n="n"/></l>
       <r>janji<s n="n"/></r></p></e>
<e><p><l>ini<s n="det"/></l>
       <r>ini<s n="det"/></r></p></e>
<e><p><l>tidak<s n="adv"/></l>
       <r>tidak<s n="adv"/></r></p></e>
<e><p><l>hak<s n="n"/></l>
       <r>hak<s n="n"/></r></p></e>
```

Figure 4. Bilingual Dictionary Entries

The bilingual dictionary records the lemma and the necessary tags such as POS tag. Compound words are recorded as one entry, for example the word "*ibu kota*" which translated as capital city, will be mapped to "*ibu negara*" (which in Indonesian will be misinterpreted as 'first lady').

```
<e><p><l>ibu~kota<s n="n"/></l>
       <r>ibu~negara<s n="n"/>
       </r></p></e>
```

Figure 5. Bilingual Dictionary Entries – Compound words

A preprocess is conducted to add tilde '~' character to combine the compound words together so that Foma will handle it as single word. This is because currently Foma does not tokenize the sentence while doing the analysis which is a functionality that other Apertium morphological tools have, such as Lttoolbox and HFST.

VII. Language Resources

In the analysis and generation step, monolingual dictionaries on both languages are needed. To build the Indonesian monolingual dictionary, we take the list of lemmata that was available before on the previous Morphological Analyser [4] and adapt it with the current setting. We keep only the lemmata that are tagged as Noun, Verb, and Adjectives. Additionally, closed word entries such as prepositions or conjunctions are added and tagged. The problem in Malaysian side is that we do not have list of Malaysian lemmata as we have in Indonesian side. We simply take the Malaysian entry on the bilingual dictionary.

Indonesian and Malaysian dictionary is not yet available. To build a fast and cheap bilingual dictionary, we grabbed available public online dictionary and also generating it from a parallel corpus. Here describes the process of the dictionary construction:

1) *Online Dictionary.* There are several online dictionary website available. We query the site for each Indonesian lemma and grabbed the translation word if available. The source tag and the target tag are also recorded.

2) *Statistical word pairing.* Word pairs are also build by using statistical method. This is done by training a small size of parallel corpus composed from several sources such as manuals, recipes, agreements, and holy books. The tools used is Moses (http://www.statmt.org/moses/) [18]. On the source language side, the words are being analyzed to get the analysis forms (lemma and morphological tags) while the target side composed of sentences with words in surface forms. After we got the word pairs, the words morphems on the target side are stripped. This is done to get lemma-to-lemma pairs.

The results from both approaches are merged and handpicked to retain the quality of the translation.

VIII. Conclusions and Future Work

Although the experiment described in the paper is still work in progress and we are at the current stage unable to provide a standard quality evaluation, there are already some results which may turn out to be important for further research.

First of all, the work on the system has led us to the investigation of both languages in the direction of how certain phenomena may be handled from the point of view of machine translation, which phenomena may cause problems in a relatively straightforward system etc.

Second, the relatively high numbers of resources needed for building individual modules for the system made us think about the methods how to obtain them in a reasonable quantity and quality. This turned out to be a challenge especially because for the European languages used in previous experiments there are many more resources available, nothing is usually built from scratch. Building better resources will be the first task which probably will help us to improve the system in the future. The development of building the full pipeline of the system didn't take most of the development time if compared to the effort on developing the resources such as morphological analyser and dictionaries.

It will be an interesting research to build the MT system in the opposite direction, Malaysian to Indonesian, which appears to be somehow symmetrical. Another challenging research would be to make Indonesian/Malaysian-English MT system using this approach.

Acknowledgment

Thanks to Måns Huldén for his help in converting patent-encumbered and some other aspects of the Xerox syntax into Foma. Thanks to Francis Tyers for the support and his help setting up the new Apertium language pair development environment.

References

[1] A. M. Corbi-Bellot, M. L. Forcada, S. Ortiz-Rojas, J. A. Prez-Ortiz, G. Ramirez-Sanchez, F. Sanchez-Martinez, I. Alegria, A. Mayor, and K. Sarasola, "An open-source shallow-transfer machine translation engine for the romance languages of spain," Proceedings of the Tenth Conference of the European Association for Machine Translation, pp. 79–86, May 2005.

[2] F. M. Tyers, F. Sánchez-Martínez, S. Ortiz-Rojas, and M. L. Forcada, "Free/open-source resources in the Apertium platform for machine translation research and development," The Prague Bulletin of Mathematical Linguistics No. 93, pp. 67-76, 2010.

[3] F. M. Tyers, L. Wiechetek, and T. Trosterud, "Developing prototypes for machine translation between two Sámi languages," Proceedings of the 13th Annual Conference of the European Association ofMachine Translation, EAMT09, 2009.

[4] F. Pisceldo, R. Mahendra, R. Manurung, and I W. Arka, "A Two-Level Morphological Analyser for Indonesian," Abstract submitted to the Australasian Language Technology (ALTA) Workshop 2008, Tasmania, 2008.

[5] F. Sánchez-Martínez, J. A. Pérez-Ortiz, and M. L. Forcada, "Using target-language information to train part-of-speech taggers for machine translation," Machine Translation, volume 22, numbers 1-2, pp.29-66.

[6] H. Dyvik, "Exploiting structural similarities in machine translation," Computers and Humanities 28, pp. 225–245, 1995.

[7] J. Hajič, J. Hric, and V. Kuboň, "Machine translation of very close languages," Proceedings of the 6th Applied Natural Language Processing Conference, 2000.

[8] J. Hajič, P. Homola, and V. Kuboň, "A simple multilingual machine translation system," Proceedings of the MT Summit IX, New Orleans, 2003.

[9] J. Vičič, "Rapid development of data for shallow transfer rbmt translation systems for highly inflective languages," Jezikovne tehnologije, language technologies : zbornik konference : proceedings of the conference, pp. 98–103, 2008.

[10] K. Altintas and I. Cicekli, "A machine translation system between a pair of closely related languages," Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002), 2002.

[11] K. Oliva, "A Parser for Czech Implemented in Systems Q," Explizite Beschreibung der Sprache und automatische Textbearbeitung XVI, MFF UK Prague, 1989.

[12] K. P. Scanell, "Machine translation for closely related language pairs," Unknown, 2008.

[13] K. Unhammer and T. Trosterud, "Reuse of free resources in machine translation between Nynorsk and Bokmål," Proceedings of the First

International Workshop on Free/Open-Source Rule-Based Machine Translation / Edited by J. A. Pérez-Ortiz, F. Sánchez-Martínez, F. M. Tyers, pp. 35-42, Alicante : Universidad de Alicante, Departamento de Lenguajes y Sistemas Informáticos, 2009.

[14] M. Hulden, “Foma: a finite-state compiler and library,” Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session, pp. 29-32, Athens, Greece, April 03-03, 2009.

[15] M. L. Forcada, F. M. Tyers, and G. Ramírez-Sánchez, “The free/opensource machine translation platform Apertium: Five years on,” Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation FreeRBMT'09, pp. 3-10, November 2009.

[16] P. Homola and V. Kuboň, “A translation model for languages of acceding countries,” Proceedings of the IX EAMT Workshop, La Valetta, University of Malta, 2004.

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

[18] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, “Moses: Open Source Toolkit for Statistical Machine Translation,” Annual Meeting of the Association for Computational Linguistics (ACL): Demonstration session, Prague, Czech Republic, June 2007.

[19] S. Marinov, “Structural Similarities in MT: A Bulgarian-Polish case,” unknown, 2003.