Persian to English Translation Problems of Topicalization Process in Apertium Platform[1]

Parya Razmdideh[2], Abbas Ali Ahangar[3],
Seyyed Mojtaba Sabbagh Jafari[4] & Gholamreza Haffari[5]

Abstract

Machine translation encounters several problems in translating from Persian language to English, due to morphological, lexical, and structural divergences between these languages. It becomes especially more difficult when the source language (SL) has specific characteristics which are unavoidable in the process of machine translation systems. This article is going to present some syntactic problems, the Apertium shallow-transfer rule-based machine translation (RBMT) platform encounters in translating structures with topilcalization from Persian to English, and tries to solve them based on the Apertium structural transfer module. Then, this developed Apertium system is evaluated using word error rate (WER) and position-independent error rate (PER), metrics and its quality is compared with that of Google translate as a statistical machine translation system. The Apertium Persian monolingual dictionary was extracted from the frequent words of Wikipedia Persian Monolingual Corpus and Persian side of Mizan English-Persian Parallel Corpus. The result shows that the syntactic translation problems mainly arise from Persian syntactic structures with topicalized constituents which are difficult to be handled by the Apertium structural transfer module. One way to solve them is writing new structural transfer rules to translate these structures more adequately.

Keywords: Apertium, RBMT, topicalization, structural transfer rule.

[1] This paper was received on 14.03.2017 and approved on 12.06.2017.

[2] Ph.D. Candidate of Linguistics, University of Sistan and Baluchestan, Iran; email: p.razmdide@gmail.com

[3] Corresponding Author: Associate Professor of Linguistics, University of Sistan and Baluchestan, Iran; email: ahangar@english.usb.ac.ir

[4] Assistant Professor of Computer Engineering, Vali-e-Asr University of Rafsanjan, Iran; email: mojtaba.sabbagh@vru.ac.ir

[5] Lecturer at Faculty of Information Technology, Monash University, Australia; email: gholamreza.haffari@monash.edu

1. Introduction

Although both Persian and English languages belong to the Indo-European language family, they differ in their phonology, morphology, and syntactic structures. These differences, especially in Persian texts, give rise to some problems in natural language processing, particularly in machine translation. Some of the syntactic structures which are different in two languages, among others, are as follows: all Persian non-verbal categories are head initial; head-complement parameter in this language is controversial, as far as verb-clausal complement order is concerned. Some believe Persian verb phrase with a complement clause is head-initial (Browning and Karimi, 1990), but Karimi (1989) and Darzi (1996) argue that it is a head-final category. Contrary to that, English is a head-initial language (Cook, 1988). In Persian, usually no article is used before nouns, whereas, English singular nouns obligatorily take a(n) (in)definite article, and plural nouns optionally appear with a definite one. Persian verbs conjugate in their own tense while a rich agreement system cannot be found in English morphology. However, "only third person singular forms of verb take any special endings" (Carnie, 2013: 414). In addition, Persian is a pro-drop language unlike English that does not have *pro* (Simpson, 2005), it allows a null subject (Dalili, 2009: 84). In this case, "the agreement (person and number) embedded in the verb can play the subject role" (Shamsfard, 2011: 65). Also, Persian and English word order has two major differences. First, English basically follows the SVO (subject-verb-object) while Persian follows, in most cases, the SOV word order. Second, English has a rigid word order but Persian allows for free word order. In Persian, the basic word order is SOV, but all of the other orders are also correct (Dabir-Moghaddam, 2001).

This study, as to the authors' knowledge, is the first research to investigate the syntactic translation problems of topicalization process between Persian and

English Apertium shallow-transfer RBMT[1] system. Apertium was firstly developed between Persian and English by authors of this article[2].

2. Apertium platform and its modules

Apertium is a shallow-transfer and free/open source[3] machine translation system which is published by developers at Alicante University consistent with GNU GPL (general public license) conditions (Forcada, Ginestí-Rosell, Nordflk, et al, 2011). Shallow-transfer machine Translation (MT) system is one of the transfer-based systems which are performed in three steps: an analysis of SL text into an SL intermediate representation (IR); transferring the IR to the target language (TL); then, generating translation from the TL intermediate representation (De(trez, Sánchez-Cartagena and Ranta, 2014).

Apertium has recently been used for the development of several language pairs (47 language pairs[4]) (http://www.apertium.org), such as Italian–Catalan (Toral, Ginestı(-Rosell and Tyers, 2011) or Icelandic–English (Brandt, Loftsson, Sigurpo(rsson, et al, 2011). It requires two types of linguistic data: dictionaries (monolingual and bilingual) and rules (structural transfer and lexical selection). They are mainly XML (Extensible Markup Language)[5]-format and hand-written. Persian monolingual entries were extracted from the frequent words of Wikipedia Persian Monolingual Corpus[6] and Persian side of Mizan English-Persian Parallel Corpus (Corpus Supreme Council of Information and Communication Technology, 2013).

[1] Part of this research was conducted at Monash University, Australia (2016-2017).

[2] https://svn.code.sf.net/p/apertium/svn/incubator/apertium-pes-eng

[3] Free Open Source Software (FOSS), also called just Open Source or Free Software, is licensed to be free to use, modify, and distribute.

[4] http://wiki.apertium.org/wiki/Language_pair

[5] http://www.w3.org/XML/

[6] fawiki-20140802-corpus.xml.bz2

Apertium platform is made of several modules to perform transfer, as represented in Figure 1:

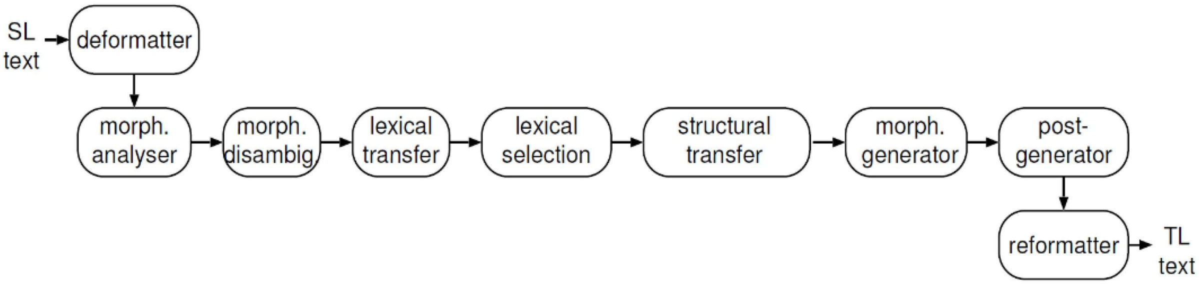

Figure 1. The pipeline architecture of Apertium system (Karibayeva, 2015)

Deformatter is the first module. Some formatting tags as HTML (Hyper Text Markup Language) are included in the source text to be separated by the deformator. These tags are called 'superblanks' to insert the space between the words in such a way that the remaining modules see them as regular blanks (Sundetova, Forcada and Tyers, 2015).

Morphological analyser is generated by compiling a morphological dictionary of SL, based on the direction it is read by the system (left to right). The morphological analyser tokenizes the text in surface forms (SF) (lexical units as they appear in texts) and delivers, for each SF, one or more lexical forms (LF) consisting of three parts: lemma (the base form commonly used in classic dictionary entries) , the lexical category or part of speech (PoS), and their morphological inflection attributes (Forcada, Bonev, Ortiz-Rojas, Ortiz, et al, 2010: 7).

PoS tagger is based on Hidden Markov Model (HHM). Every SL surface form should be analyzed to one TL lexical form. Ambiguous words in SL morphological dictionary are analyzed to more than one TL lexical form. To solve this problem, the PoS tagger module selects one of these lexical forms, and sometimes constraint grammar rules (Karlsson, 1990) are applied before the final result of PoS tagger.

The operation of **Lexical transfer** module is based on a bilingual dictionary. It maps SL lexical forms to a TL lexical form from the bilingual dictionary. Some translation ambiguities can be handled using multi-word expressions (MWEs) encoded in the dictionaries of the system, but the status quo is that, for any given SL word, the most frequent, or most general translation is given. This poses a translation problem, as often it may be difficult to choose the most frequent, or the selection strongly depends on the context (Tyers, Sánchez-Martínez and Forcada, 2012). To solve this lexical ambiguity, the next module is used.

Lexical selection rules can be written manually to solve the problem of lexical ambiguity. If there are more than one TL equivalent for one SL lexical form, this module selects one of them.

Structural transfer module contains three levels of transfer rules which are written manually. Structural transfer in Apertium applies a set of rules in a left-to-right, longest-match fashion to prevent word for word translation in those cases in which this would result in an incorrect translation (Forcada, Bonev, Ortiz-Rojas, Ortiz, et al, 2010).[1]

Structural transfer rules detect and process patterns of words (chunks or phrases) based on grammatical divergence between the two languages. They are split into three levels to facilitate the writing of rules by linguists (Forcada, Bonev, Ortiz-Rojas, Ortiz, et al, 2010). The first level is called **chunk**, performing short-distance operations into *chunks;* The second-level rules are called **interchunk**, performing *interchunk* operations, like agreements between more distant constituents, and third-level is called **post-chunk**, de-encapsulating the chunks and generating a sequence of TL lexical forms from each *chunk* (Sa(nchez-Cartagena, Sánchez-Martínez and Pe□rez-Ortiz, 2011).

[1] The phrase داستان زیبا را will be detected and processed by the rule for 'noun- adjective-postposition' not by the rule for 'noun-adjective'.

Persian to English Apertium system is summarized in Table 1.

Table 1. The current status of Persian to English Apertium system

Monolingual Persian Dictionary	18285 lemmas[1]
Bilingual Persian to English Dictionary	6856 lemmas
Monolingual English Dictionary[2]	87481 lemmas
Persian to English Lexicon Selection Rules	78 rules
Persian to English Constrain Grammar Rules	10 rules
Structural Transfer rules (apertium-pes-eng.pes-eng.t1x)	157 rules
Structural Transfer rules (apertium-pes-eng.pes-eng.t2x)	23 rules

Morphological generator is a letter transducer that generates a surface form in the TL from the bilingual transfer output (Peradin and Tyers, 2012).

In **Post-generator**, some minor orthographical changes are performed at TL lexical forms (Karibayeva, 2015). Finally, **Reformattor** backs the formatting tags to the text to make it look as it appeared first (Trosterud and Unhammer, 2012).

3. Related Works

There are many works done on Apertium platform. But this article is the first effort to consider translation problems of topicalized constituents based on Persian to English Apertium platform. Here, some of the most relevant studies are considered.

Tyers and Nordfalk (2009) described the development of a shallow-transfer RBMT system between Swedish and Danish. As these languages are both Germanic, there are few translation problems on the syntactic level. Then, the system was evaluated quantitatively using WER and PWER (position-independent word error rate) and some of its translation problems were discussed. Varga and Yokoyama

[1] The number of lemmas in Persian monolingual dictionary is more than their corresponding in bilingual dictionary mostly because of Persian verbs which can be written in different orthographical ways.

[2] We used the existing English monolingual dictionary between English and Kazakh language pair (Sundetova, Forcada and Tyers, 2015).

(2009) attempted to generate the most frequent transfer rules using a small or medium sized parallel corpus and a bilingual dictionary, concentrating mainly on word-level and inflectional correspondences. The results displayed that with a small corpus or medium sized corpus with noisy parsers, the grammatical exceptions or idiomatic expressions weren't handled. Forcada, Bonev, Ortiz-Rojas, Ortiz, et al (2010) is the first introductory document that describes Apertium modules and how to develop, install, and run an Apertium system for a new language pair. Tyers, Sánchez-Martínez, and Forcada (2012) described a model to process lexical selection rules to overcome the lexical selection problems in RBMT system. Based on the results, in pair bootstrap resampling, the system offered a statistically significant improvement in translation quality over the next highest scoring system. Amirova (2015) implemented a statistical method, maximum entropy model, to solve the problem of lexical selection rules for English-Kazakh language pair in Apertium. The results showed that by 85 lexical selection rules in English-Kazakh, the system could translate simple phrases and sentences with ambiguity. De(trez, Sánchez-Cartagena, and Ranta (2014) developed two methods to share linguistic data between Apertium and Grammatical Framework (GF). Two sharing strategies were augmenting the GF lexicon with Apertium dictionary entries and inferring Apertium shallow transfer rules from GF grammars. The results indicated that an Apertium-based system was created without manually witting a single-shallow transfer rule and a GF-based system outperformed Apertium word for word translation on the smallest corpus.

4. Translation Problems of Topicalized constituents in Apertium

To implement structural transfer module in Apertium from Persian to English, some syntactic problems are considered mostly as a result of Persian syntactic structures. Persian has a canonical SOV word order; however, there are also lots of frequent exceptions in word order, caused by processes such as topicalization, dislocation, clefting, pseudoclefting, and scrambling that mainly result in high structural complexity (Saedi, 2009). This article investigates translating challenges of Persian topicalization process by Apertium and solves them by adding new structural transfer rules at *apertium-pes-eng.pes-eng.t1x* file[1]. All examples were used from both basic sentences and non-basic sentences (For information on these sentences see: Yarmohammadi, 2012: 33–34).

4.1. Topicalization

The most common way to topicalize an element is to move it to the initial position of a sentence. Topicalization is optional and is used for all phrases (Mahootian, 1997). Similarly, this process can occur in both main and subordinate clauses.

4.1.1. Topicalization in the main clause

In a main clause, noun phrases can be topicalized as subjects, direct objects, indirect objects, etc. Topicalization can be applied to them via the use of the postposition 'ra'. The subject and the verb don't take 'ra' for topicalization (Mahootian, 1997).

4.1.1.1. Topicalized subject

	Topicalize d subject	Initial Apertium Translatio n	Apertium structural transfer rule(s)	Final Apertium translatio n
۱. آیا هوا سرد بود؟ Was it	۲. هوا، آیا سرد بود؟ Was it cold?	weather was cold was?	هوا[۲]<n>/weather<n> آیا<vrbser>/was<vblex> سرد<adj>/cold<adj>بود<vblex>/was<vblex> ?[3] Rule 2 was<vblex> the <det><def> weather<n> cold<adj>?	[4]was the weather cold?
	Persian pattern[5]			English pattern

[1] It can be downloaded from https://svn.code.sf.net/p/apertium/svn/incubator/apertium-pes-eng.

[2] All Persian lexical forms are generated based on Morphological Aanlyser module.

[3] The symbols are used from http://wiki.apertium.org/wiki/List_of_symbols.

[4] Currently the system is not sensitive to capital letters at the beginning of a sentence.

[5] All Persian and English patterns are based on the morphological attributes of lemmas and their orders at structural transfer rules.

cold ?	n-vblex-adj-vblex?			vblex-the-n-adj?

The attributive sentence[1] (2) with topicalized subject included the patterns: noun '<n>', main verb ('to be') '<vblex>', adjective '<adj>', and main verb '<vblex>'. Rule (2) reordered them to the main verb-the-noun-adjective?

4.1.1.2. Topicalized direct objects encompass generic, definite, and indefinite direct objects (Mahootian, 1997).

a) Topicalized generic direct object

	Topicalized subject	First Apertium Translation	Apertium structural transfer rule(s)	Final Apertium translation
۳. نباید ماهی بخری. You shouldn't buy fish	۴. ماهی، نباید بخری. Fish, you shouldn't buy.	fish shouldn't buy.	ماهی<n>/fish<n> Rule 150↓ fish<n> نباید<vaux><inf><neg>/should<vaux><inf><neg> بخری<vblex><pres>/buy<vblex><pres> Rule 30↓ you<prn><subj> should<vaux><inf> n't<adv> buy<vblex><pres>	fish you shouldn't buy.
	Persian pattern			English pattern
	n-vaux(inf)-vblex			n-prn-vaux(inf)-vblex

To generate the topicalized generic direct object in sentence (4), rule (150) produced the noun 'fish' at the beginning of the sentence. Then, the rest of the sentence was translated by rule (30), including the subjective pronoun 'you', the negative modal verb 'shouldn't', and the infinitive '<inf>' verb 'buy'.

b) Topicalized definite direct object

	Topicalized	Initial	Apertium structural transfer rule(s)	Final

[1] To generate non-attributive sentences with topicalized subject as "رضا، آیا خوابید؟" the same rules were used.

	definite direct object	Apertium Translation		Apertium translation
۵. بچه غذا را خورد. The child ate the food.	۶. غذا را، بچه خورد. The food, the child ate.	food child eat.	غذا<n>/food<n> را[1]<det><def>/the<det><def> بچه <n>/child<n> خورد<vblex><past>/ate<vblex><past> Rule 18↓ the<det><def> food<n> the<det><def> child<n> eat<vblex><past>	the food the child ate.
	Persian pattern			English pattern
	n-det(def)-n-vrb			the-n-the-n-vrb

To topicalize a definite direct object in sentence (6), rule (18) changed the noun "غذا" and object marker "را" to noun phrase 'the food'. Also the noun "بچه" and the verb "خورد" were translated to 'the child' as a noun phrase and the past tense '<past>' verb '<vblex>' 'ate'.

c) Topicalized indefinite direct object

	Topicalized definite direct object	Initial Apertium Translation	Apertium structural transfer rule(s)	Final Apertium translation
۷. رضا یک کتاب خرید[2]. Reza bought a book.	۸. یک کتاب را ، رضا خرید. A book, Reza bought.	one book Reza bought.	یک<det><ind>/a<det><ind> کتاب<n>/book<n> را<det><det>/the<det><def> Rule 140↓ a<det><ind> book<n> رضا<np>/Reza<np>	a book Reza bought.
	Persian			English

[1] The postposition 'ra' in Persian was matched with 'the' as definite determiner in English. Based on Apertium two lemmas should have the same or similar morphology to be equivalent with each other (Forcada, Bonev, Ortiz-Rojas, Ortiz, et al, 2010).

[2] The same rules were applied to generate the similar sentences like "کتابی را، رضا خرید" or "یک کتاب را، رضا خرید".

	pattern		Rule 150↓ Reza<np> خرید<vblex><past>/buy<vblex><past> Rule 34↓ buy<vblex><past>	pattern
	det(ind)-n-det(def)-np-vblex			det(ind)-n-np-vblex

Sentence (8) with topicalized indefinite direct object was translated by three rules. They were matched from left to right to topicalize indefinite direct object to the initial position. The rule (140) produced the indefinite noun phrase "یک کتاب" with 'a' as indefinite '<ind>' determiner '<det>' and the noun phrase 'a book'. Rule (150) generated the proper noun '<np> "رضا". Finally, the past tense verb "خرید" was generated as 'bought' by rule (34).

4.1.1.3. Topicalized indirect object

	Topicalized indirect object	Initial Apertium translation	Apertium structural transfer rule(s)	Final Apertium translation
۹. رضا . کتاب را به حسن داد. Reza gave the book to Hasan.	۱۰.حسن را، رضا کتاب را بهش[1] داد. Hasan, Reza gave the book to.	to Hasan Reza book gave.	حسن<np>/Hasan<np> را<det><def>/the<det><def> Rule 120↓ Hasan<np> رضا<np>/Reza<np> کتاب<n>/book<n> را<det><def>/the<det><def> بهش<pr><enc><pos>/to <pr><enc<<pos> داد<vblex><past> /give<vblex><past> Rule 61↓ Reza<np> give<vblex><past> the<det><def> book<n> to<pr>	Hasan Reza gave the book to.
	Persian pattern			English pattern
	np-det(def)-np-n-det(def)-pr(enc)-vrb			np-np-vrb-the-n-pr

[1] The resumptive pronouns maybe used either as personal pronouns "... من، تو، او" or enclitic pronouns "... ـِم، ـِت، ـِش".

Rule (120) topicalized the indirect object "حسن را" to the beginning of the sentence (10) and placed an enclitic pronoun "-ِش" in its original position. Then rule (61) translated the proper noun "رضا", the past tense verb "داد", the definite noun phrase "کتاب را", and the enclitic pronoun with preposition "بهش" into 'Reza gave the book to'.

4.1.2. Topicalization in the subordinate clause

Noun phrases as subjects, direct, and indirect objects in a subordinate clause can also be topicalized to sentence-initial position (Mahootian, 1997). Topicalizing every constituent except for the subject in subordinate clauses can be used via the use of 'ra' (Dabir-Moghaddam, 1990). The following sentences illustrate the topicalized subject, direct object, and indirect object in the subordinate clause.

4.1.2.1. Topicalized subject

	Topicalized subject	Initial Apertium translation	Apertium structural transfer rule(s)	Final Apertium translation
۱۱. (من) گفتم که علی کتاب را به حسن بدهد. I told Ali to give the book to Hasan.	۱۲. علی را، (من) گفتم که کتاب را به حسن بدهد. Ali, I told to give the book to Hasan.	Ali I told that book to Hasan gave.	علی<np>/Ali<np> را<det><def>/the<det><def> گفت<vblex><past>/tell<vblex><past> Rule 56↓ Ali<np> I<prn><subj> tell<vblex><past> کتاب<n>/book<n> را<det><def>/the<det><def> به<pr>/to <pr> حسن<np>/Hasan<np> ده<prs>/give<prs> Rule 57↓ to <pr> give<prs> the<det><def> book<n> to<pr> Hasan<np>	Ali I told to give the book to Hasan.
	Persian pattern			English pattern
	np-det(def)-(prn)-vblex-rel-n-det(def)-pr-np-vblex			np-prn-vblex- to-vblex-the-n-pr-np

Rules (56) and (57) were applied to translate sentence (12) with topicalized subject. They are summarized as below:

proper noun+ definite determiner + (subjective pronoun) + verb

Rule 56↓

proper noun +subjective pronoun + verb

In sentence (12), the proper noun "علی" was topicalized to the beginning of the sentence. Then the subjective pronoun 'I' and the past tense verb 'told' were written in its translation. Rule (57) produced the rest of the sentence:

noun + definite determiner + preposition + noun + verb

Rule 57↓

to + verb + 'the' + noun + preposition + noun/proper noun

The infinitive marker 'to' was inserted before the infinitive verb 'give'. As 'book' is a singular countable common noun, the definite determiner 'the' was used before it. Then the prepositional phrase 'to Hasan' was added to the end of the sentence.

4.1.2.2. Topicalized direct object

	Topicalized direct object	Initial Apertium translation	Apertium structural transfer rule(s)	Final Apertium translation
۱۳. (من) گفتم که علی کتاب را به حسن بدهد. I told Ali to give the book to Hasan.	۱۴. کتاب را، (من) گفتم که علی آن را به حسن بدهد. The book, I told Ali to give to Hasan.	book I told that Ali it to Hasan gave.	کتاب<np>/book<np> را<det><def>/the<det><def> گفت<vblex><past>/tell<vblex><past> Rule 56↓ the<det><def> book<n> I<prn><subj> tell<vblex><past> علی<np>/Ali<np> آن<prn><obj>/it<prn><obj> را<det><def>/the<det><def> به<pr>/to <pr> حسن<np>/Hasan<np> ده<prs>/give <prs> Rule 59↓ Ali<np> to <pr> give<pres> to<pr>	the book I told Ali to give to Hasan.
	Persian pattern			English pattern
	n-det(def)-(prn)-vblex-rel-			def(def)-n-prn-vblex-np-to-

	np-prn-det(def)-pr-np-vblex		Hasan<np>	vblex-to-np

The first part of the sentence (14) "کتاب را (من) گفتم" was generated following rule (56) as it was used in sentence (12). Then rule (59) translated the present subjunctive verb "بدهد" into infinitive 'to give' and omitted the resumptive pronoun "آن" in English pattern.

4.1.2.3. Topicalized indirect object

	Topicalized indirect object	Initial Apertium translation	Apertium structural transfer rule(s)	Final Apertium translation
۱۵. (من) گفتم که علی کتاب را به حسن بدهد. I told Ali to give the book to Hasan.	۱۶. حسن را، (من) گفتم که علی کتاب را بهش بدهد. Hasan, I told Ali to give the book to.	Hasan I told that Ali book to him gave.	حسن<np>/Hasan<np> را<det><def>/the<det><def> گفت<vblex><past>/tell<vblex><past> Rule 56↓ Hasan<np> I<prn><subj> tell<vblex><past> علی<np>/Ali<np> Rule 150↓ Ali<np> کتاب<n>/book<n> را<det><def>/the<det><def> بهش<pr><enc><pos>/to <pr><enc<<pos> ده<prs>/give<pres> Rule 60↓ Ali<np> give<pres> the<det><def> book<n> to<pr>	Hasan I told Ali *gives the book to.
	Persian pattern			English pattern
	np-det(def)-(prn)-vblex-rel-np-n-det(def)-pr(enc)-vblex			np-prn-vblex-np-vblex-the-n-pr

In generating indirect object in sentence (16) three rules were applied. Rule (56) was used the same as in sentence (12). Rule (150) produced the proper noun 'Ali'. Then rule (60) translated the patterns as it was used in sentence (10). But the problem is that 'to', as the infinitive marker, should be inserted before bare present verb to produce an infinitive form with 'to' as an infinitive marker, a case which needs more investigation.

5. Evaluation

To evaluate Persian to English Apertium system, two evaluation metrics were used: WER and PER. WER calculates the minimum number of substitutions, deletions, and insertions which are performed to convert the generated text into the reference text. The WER shortcoming is that it doesn't allow reordering of words, whereas the word order of the generated text is different from that of the reference text even it is a correct translation meaningfully. To solve this problem, PER of two sentences is compared without taking the word order into account. The PER is always lower than or equal to the WER (Popovič and Ney, 2007).

To make a quantitative evaluation, 100 Persian sentences were extracted from Mizan English-Persian Parallel Corpus which mostly contained different types of syntactic structures with topicalized constituents. The sentences were translated by Apertium as a RBMT system and Google as a statistical machine translation system. Then they were post-edited by a human translator. Both machine translated sentences were evaluated by WER and PER[1] in the developed system and Google Translate.

The results of quantitative evaluation of WER and PER in Apertium and their comparison with Google are shown in table 2. It indicates that Apertium output is closer to the reference translation than that of Google translate.

Table 2. Comparative evaluation results for post-edition task

Machine Translation System	WER	PER
Apertium	56.64 %	44.25 %
Google	62.78 %	42.02 %

[1] To download the freely available tool from apertium-eval-translator, refer to http://apertium.org

Based on the evaluation results given in table (2), the distance between WER scores (6.14%) in both Apertium and Google Translate is more than their PER scores (2.23%). The reason is that WER takes into account the word order in contrast to PER which does not penalize word order in the translation (Costa-Jossa`, 2012). So, adding any structural transfer rule with the most effect on word order covering the translated sentences with the topicalized constituents decreases WER score at Apertium platform.

6. Conclusion

This article considered the translation problems of topicalization process between Persian and English in the free/open-source Apertium platform. All examples included basic and non-basic sentences. The topicalized constituents were from both main and subordinate clauses such as subjects, direct objects, and indirect objects. These problems were mostly a result of Persian syntactic structures. To overcome them, some new structural transfer rules were written for *apertium-pes-eng.pes-eng.t1x*. To evaluate the system, 100 sentences were extracted from Mizan English-Persian Parallel Corpus which mostly had structures with topicalized constituents. These sentences were translated by Apertium and Google translate. Comparing WER and PER evaluation metrics in Apertium and google showed that Apertium could translate these syntactic challenges more adequately than Google. In terms of future works, we intend to cover other Persian syntactic structures like clefting, pseudoclefting, and extraposition. Furthermore, we are in the process of improving the system coverage in lexicon and especially in structural transfer rules with a machine learning method, named Active Learning (AL).

Acknowledgements

We are highly grateful to Francis Tyers and Víctor Manuel Sánchez Cartagena for answering our questions during the work.

Works Cited:

Amirova, D. (2015). Choosing the model for solving the problem of lexical selection for English-Kazakh language pair in the free/open-source platform Apertium. *3rd International Conference on Computer Processing in Turkic Languages.*

Brandt, M. D., Loftsson, H., Sigurpo(rsson, H., & Tyers, F. M. (2011) .Apertium-Ice NLP: a rule-based Icelandic to English machine translation system. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, 217–224.

Browning, M., Karimi, E. (1990). Scrambling in Persian. *Tiburg Scrambling Conference.*

Carnie, A. (2013). *Syntax: A Generative Introduction.* Blackwell Publishing, Second Edition.

Cook, V. J. (1985). Universal Grammar and Second Language Learning. *Applied Linguistics*, 6, 2–18.

Costa-Jossa`, M., R. (2012). Study and Comparison of Rule-based and Statistical Catalan-Spanish Machine Translation Systems. *Computing and Informatics.* 3, 245–270.

Dabir-Moghaddam, M. (2001). Word Order Typology of Iranian Languages. *Journal of Humanities*, 8(2), 17–24.

Dalili, V. M. (2009). Agreement (AGR) and the Pro-drop/Non-pro-drop Variation: A Meta-analysis of GB and MP accounts. *Philologie im Netz,* 49, 84–102.

Darzi, A. (1996). *Word Order, Np-Movement, and Opacity Conditions in Persian.* Ph.D. Dissertation, University of Illinois, Urbana.

Détrez, G., Sánchez-Cartagena, V. M., & Ranta, A. (2014). Sharing resources between free/open-source rule-based machine translation systems: Grammatical Framework and Apertium. *9th International Conference on Language Resources and Evaluation*, 4394–4400.

Farshidvard, Kh. (2013). *Contemporary Detailed Grammar.* Scientific Publications .Fourth Edition.

Forcada, M. L., Bonev, B. I., Ortiz-Rojas, S., Ortiz, J. A. P., Sánchez. G. R., Sánchez-Martínez, F., Armentano-Plller. C., Montava, M.A., & Tyerz, F. M. (2010). *Documentation of the open-Source Shallow-Transfer Mschine translation Platform Apertium.* Departament de Llenguatges i Sistemes Inform`atics Universitat d'Alacant

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., & Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. Machine translation, 25(2):127–144.

Karibayeva, A. (2015). Lexical selection rules for Kazakh-to-English machine translation in the free/open-source platform Apertium. *3rd International Conference on Computer Processing in Turkic Languages.*

Karimi, S. (1989). Aspects of Persian Syntax, and the Theory of Government. Ph.D. Dissertation. University of Washington.

Karlsson, F. (1990). Constraint Grammar as a Framework for Parsing Running Text. *In H. Karlgren, ed., Proceedings of the 13th Conference on Computational Linguistics.* Helsinki: Finland, 3, 168–173.

Mahootian, Sh. (1997). *Persian. (Descriptive Grammars).* London: Routledge.

Peradin, H., Tyers, F. (2012). A rule-based machine translation system from Serbo-Croatian to Macedonian. *In Proceedings of a workshop on Free/open source machine translation.* Gothenburg, 41–55.

Perry, J. R. (2005). A Tajik Persian reference grammar. *Brill Academic Pub,* 11.

Popovi. M., Ney. H. (2007). Word Error Rates Decomposition over PoS Classes and Applications for Error Analysis. *Association for Computational Linguistics.* Prague: Proceedings of the Second Workshop on Statistical Machine Translation, 48–55.

Radford, A. (2009). *Analysing English Sentences: A Minimalistic Approach*. Cambridge University Press.

Saedi, Ch., Shamsfard, M., & Motazedi, Y. (2009). Automatic Translation between English and Persian texts. *In Proceedings of the Third Workshop on Computational Linguistics*.

Sánchez –Cartagena, V. M., Sánchez-Martínez, F., & Pe□rez-Ortiz., J. A. (2011). The Universitat d'Alacant hybrid machine translation system for WMT 2011. *Association for Computational Linguistics*. Proceedings of the 6th Workshop on Statistical Machine Translation. Edinburgh, 457–463.

Sánchez-Martínez, F., Ramírez-Sánchez, G., & Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. Machine translation, 25(2):127–144.

Shamsfard, M. (2011). Challenges and Open Problems in Persian Text Processing. *Proceedings of LTC*, 65–69.

Simpson, A. (2005). Pro-drop Patterns and Analyticity. *LSA 222 Syntactic Analyticity*.

Sundetova, A., Forcada, M. L., & Tyers, F. (2015). A free/open-source machine translation system for English to Kazakh. *3rd International Conference on Computer Processing in Turkic Languages*.

Supreme Council of Information and Communication Technology (2013). *Mizan English-Persian Parallel Corpus*.Tehran, I.R. Iran, Retrieved from the website: http://dadegan.ir/catalog/mizan.

Toral, A., Ginestı́-Rosell, M., and Tyers, F. M. (2011). An Italian to Catalan RBMT system reusing data from existing language pairs. *In Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, 77–81.

Trosterud, T., Unhammer, K. B. (2012). Evaluation North Sa´mi to Norwegian Assimilation RBMT. *In Proceedings of a workshop on Free/open source machine translation,* 1–13.

Tyers, F. M & Nordfalk, J. (2009). Shallow-Transfer Rule-Based Machine Translation for Swedish to Danish. *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, 27–33.

Tyers, F. M., Sánchez-Martínez, F., & Forcada, M. L. (2012). Flexible finite-state lexical selection for rule-based machine translation. *Proceedings of the 16th EAMT Conference*, Trento, Italy.

Varga, I., Yokayama, Sh. (2009). Transfer rule generation for a Japanese-Hungarian machine translation system. *In Proceedings of the Machine Translation Summit XII*, Ottawa, Canada.

Yarmohammadi, L. (2012). *A Contrastive Analysis of Persian and English*. Payame Noor University Press.

دبیرمقدم، محمد (۱۳۶۹). پیرامون «را» در زبان فارسی. *مجلهٔ زبان‌شناسی*، ۷(۱)، ۶۰–۱.