A corpus-based translation study on English-Persian verb phrase ellipsis

Mitra Shahabi and Jorge Baptista
University of Algarve, Portugal

Abstract
The present research is a descriptive corpus-based translation study aiming at pinpointing the patterns of translation into Persian when dealing with English Verb Phrase Ellipsis (VPE). After scrutiny of the strategies applied by Persian translators some regular patterns were drawn, with the exception that the observed translation behavior may be taken as advantageous information for improving English-Persian Machine Translation (MT) system performance.

1 Introduction

Any research in translation should start with observational facts (translated utterances and their constituent elements) towards the reconstruction of non-observational facts (Toury 1994: 18). The translated texts should be studied focusing on the strategies the translators have adopted in dealing with, principally, the contrasts of the languages in question. This kind of study helps state some predictive or explanatory rules about those contrasts and helps achieve a good understanding of the probable errors due to the differences between patterns in the first language (L1) and those found in the second language (L2). Understanding the nature of those errors is necessary for resolving them.

According to Malmkjær (1998), a parallel corpus gives a comparative view of characteristics of translated texts, based on which target language norms can be identified. The identification of these norms can be exploited in different fields such as machine translation programs, bilingual dictionaries, language learning/ teaching purposes, and translator training.

The goal of the present study was to scrutinize the systematic strategies used in the translation of English VPE structures into Persian, thus helping to identify translation norms. By exploring the probable regularities to be found in these strategies, it may be possible to help define rules for improving the performance of English-Persian Machine Translation (MT) systems.

This paper is structured as follows: Section 1 is an introduction to the study; Section 2 presents a brief definition of verb phrase ellipsis, a brief overview of its study in NLP farmework, and VPE in translation studies; Section 3 provides the methodology; Section 4 is devoted to data analysis and discussion; Section 5 assesses the results; and finally, Section 6 presents the conclusion.

2 Verb Phrase Ellipsis (VPE)

In general, for a sentence to be complete it should contain a verbal constituent. However, sometimes sentences can be found that do not present an inflected verb form; yet they are intuitively complete (example (1)). In fact, repeated material can be zeroed to avoid redundancy (the part(s) in square brackets indicate the elided element(s)):

(1) *John read the magazine and Mary* [*read*] *the newspaper.*

Although there are extensive theoretical studies on VPE, to the best of our knowledge, there is apparently no considerable work on VPE in the NLP (Natural Language Processing) framework. In the context of anaphora resolution, for example, other kinds of anaphora have received much more attention and considerable work exists on their resolutions (Lappin 1996; Lappin and Elabbas 1999; Mitkov, Boguraev and Lappin 2001; Mitkov, Evans and Orasan 2002; Mitkov 2002; Lappin 2005). However, ellipsis as a "zero anaphora" (Mitkov 2002: 13), and particularly VPE or "verb phrase zero anaphora", has not benefitted from such extensive interest.

To the best of our knowledge, the only works focusing on automatically detecting verb phrase ellipses, identifying their antecedents, and resolving ambiguities are by Lappin and McCord (1990), Hardt (1997), Nielsen (2005), and De Vries (2009). However, their study did not extend to the MT field.

The main concern of this study is translation and the processes which help the improvement of MT performances. Human translators usually have no serious problem in dealing with ellipsis in source text, because they intuitively understand the meaning of elliptic sentences, so that they can recover easily the deleted material and fill in the missing words in their translation. MT systems, on the other hand, require some predefined information to be available. This information should be provided for MT systems; otherwise the resultant gaps lead to translation failure. English-Persian MT systems (e.g. Google Translator (GT)) fails to resolve VPE, as in example (2):

(2) *Did you go to the cinema yesterday?*
- *No, I didn't* [*go (to cinema yesterday*].

GT
آیا رفتن به سینما دیروز؟
Aya raftan be sinema dirooz?
INT GO/GERAND TO/PRE CINEMA/LOC YESTERDAY
Did going to cinema yesterday?

- نه، من نکردم.
Na, man nakardam
NO I/NOM NEG+DO/PST-1SG
No, not-did-I.
No, I didn´t do.

In example (2), Google translator translated the auxiliary verb *did* which led to a wrong output. A possible, appropriate translation is presented below (Persian sentences appearing in Arabic script are followed first by their transliteration, morphological analysis, literal translation, and then natural translation).

HT
دیروز سینما رفتی؟
dirooz sinema rafti
YESTERDAY CINEMA/LOC GO/PST/2SG
yesterday cinema went-you?
Did you go to the cinema yesterday?

- نه، نرفتم.
b) Na narafatm.
NO NEG+GO/PST/1SG
b) *No, not went-I.*
b) *No, I didn´t go.*

As noticed, the ellipsis has been filled by the antecedent verb نرفتم (naraftam) [NEG+GO/PST/1SG] 'not went-I' (I didn't go). In Persian, there is no ellipsis of single verbs; therefore, the antecedent verb رفتن (raftan) 'go' appears in the answer.

3 Methodology

In this section, the material, the tools, the framework on which this study was based, and the procedure are presented.

3.1 Corpus

For this study, a bilingual, unidirectional, English-to-Persian corpus (Pilevar 2010) was applied. It consisted of movie subtitles (612,086 parallel sentences: about 4 million words in each language). Although the general quality of the translated subtitles was good, some textual (mostly spelling) problems were found that hindered the detection of some instances of VPE. In order to improve the recall of the search pattern used for this task, the English text was pre-processed and some faulty spellings were normalized, such as no capitalization information, some lack of punctuation, bad or wrong spelling, and incorrect contraction forms.

3.2 Typology of English VPE

Based on Halliday and Hasan's (1976), a classification of English verbal ellipsis was obtained. This was complemented with remarks taken from Lobeck (1995), as well as Lappin and McCord (1990). This typology served as the base for building the search patterns applied to the English corpus in order to retrieve instances of VPE. These sources are briefly presented below.

According to Halliday and Hassan (1976), VPE occurs in sentences with auxiliary verbs and no main verb. Auxiliary verbs take the position of the main verb and lead into VPE; they are tense operators, modal verbs, or dummy verbs *(do*, *does*, *did*).

According to Lobeck (1995), the linguistic constructions in which English VPE usually occurs at the end of sentence are: coordinate or subordinate clauses, *yes*/*no* short responses, and comparative sentences. To the above list, the Lappin and McCord (1990) reference to the occurrence of VPE after the complementizer *to* (example (3)) was added.

(3) *I apologized to the teacher; I did not want to* [*apologize to him*].

3.3 Tools

Three software tools were used for different purposes: Unitex [1] (Paumier, 2008); Python [2]; and Google translator.

Unitex, an open-source corpus processing software, was applied in order to normalize the English corpus, build search patterns in the form of finite-state graphs, and apply them to the corpus in order to retrieve candidate VPE instances. Python was used to align the parallel sub-corpra of English, Persian, and Google Translation. Google translator was applied in order to analyze its performance in dealing with VPE. From among the free online MT systems, only three contained the English-Persian language pair: Babylon 8, Google, and SDL international. After a preliminary testing, Google translator was deemed to perform better than the other two and so it was selected for this study.

3.4 Procedure

Based on the typology of English VPE, a set of search patterns were defined; they were able to pinpoint the English VPEs occurring in sentences ending with a) auxiliary verbs, b) the infinitival complementizer *to*, and c) pro-forms. In order to simplify the task, only the simplest patterns were considered here, i.e. a clause consisting of only the essential elements, possiblly with a facultative adverb as well.

The search pattern was, then, given to Unitex to automatically detect the instances of VPE in the English corpus.

10,515 cases of VPE were captured. All the instances of type (b) and (c), namely 665 and 191 instances, respectively, were studied. For the largest type of matches (9,659 cases), 1,477 instances were extensively studied. They were restricted to certain linguistic contexts such as: *yes*/*no* short responses to questions; the clauses after coordinating conjunctions *and*/*but* and after adverbial conjunctions *after*/*before*; conditional clauses; and some comparative structures. From the rest of the output (8,182), consisting of different subordinate clauses or closely tied rejoinders, all the instances of VPE with auxiliary *have* were studied. About 50 per cent of the occurrences of VPE with modal verbs and each of the other two kinds of operators (*do* and *be*), from three random locations of the corpus were studied. The first overview of the resulting matches from the corpus is shown in Table 1:

Table 1: The first overview of the resulting matches from the corpus

VPE in clauses ending with	Number of matches		Aux.	Number of occurrences	Studied instances
auxiliary verbs	9,659	Subordinates & coordinated (1,477)	*do*	436	436
			be	675	675
			have	152	152
			modals	214	214
		Others (8,182)	*do*	2,034	1,000
			be	2,969	1,500
			have	488	488
			modals	2,691	1,300
complementizer *to*					665
pro-forms					191
Total					6,621

The sentences presenting instances of VPE provided the English sub-corpus. The Persian counterparts of the English sub-corpus provided the Persian sub-corpus. In order to improve the performance of English-Persian MT, it was necessary to detremine MT failure in treating English VPE. For this purpose the English sub-corpus was translated by Google translator, and the GT sub-corpus was produced.

In order to verify the representativeness of the results of the study, a sampling was carried out using a randomly- extracted portion of text from 6 different locations of the entire corpus (each containing 5% of the corpus, in total 183,607 sentences). This sub-corpus was then manually analyzed and the spotted instances of VPE were compared against their Persian counterparts in order to assess the translation strategies adopted for them. Naturally some patterns had already appeared in the first data analysis procedure. It should be noted that VPEs after complementizer *to*, and VPEs in pro-forms structure were few and they were all studied.

4 Data analysis and discussion

In this section, the VPE instances are analyzed. First, VPE occurring after operators (*do*, *be*, and *have*) and modal verbs are discussed; then, VPE after the infinitival complementizer *to*; and finally, VPE in pro-form structures are presented.

i. VPE with tense and modal auxiliary verbs

5,765 cases of VPE after the auxiliaries *do*, *be*, *have*, and modals were pinpointed, from which 2,071 cases were irrelevant to the study (Table 2):

Table 2: Number of VPE after auxiliary verbs (relevant & irrelevant cases)

Aux.	Studied cases	Relevant cases	Irrelevant cases
do	1,436	1,025	411
be	2,175	1,054	1,121
have	640	480	160
modals	1,514	1,135	379
Total	5,765	3,694	2,071

It was found that, in English, all verbs are subject to VPE (example (4), 0, and (6)) while, in Persian, VPE is only possible in the presence of some modal verbs (example 0) or when the verb is in the simple past/present passive voice (example (6)). Consider examples (4) to (6):

(4) *He usually talks all the time. He didn´t* [*talk all the time*] *yesterday.*

HT

اون معمولا یریز حرف میزنه. دیروز هیچ حرف نزد .

Oon mamoolan yeriz harf mizane. Dirooz hich harfi nazad.

HE/NOM USUALLY CONSTANTLY WORD/NC+HIT/LV/PRS-3SG.YESTERDAY NOTHING WORD/NC+NEG +HIT/LV/PST-3S

He usually hits_word constantly.Yesterday not-hit_word nothing.

He usually talks constantly. Yesterday he didn't talk

.

(5) *You cannot live alone forever.*
- *I can [live alone].*

HT
تو نميتوني تا ابد تنهايي زندگي كني.
To nemitooni ta abad tanhai zendegi koni.
YOU/NOM-2SG NEG+CAN-2SG TILL EVER ALONE LIFE/NC+DO/LV-2SG
You not-can till ever alone do-you_life.
You cannot live alone for ever.
- *من ميتونم [تنها زندگی کنم].*
Man mitoonam [tanha endegi konam.]
I/NOM CAN/MOD-1SG [ALONE LIFE/NC+DO/LV-1SG]
I can [*alone do-I_life*].
I can [*live alone*].

(6) *Are you bored?*
- *No, I'm not.*

HT
خسته شدی؟
Khaste shodi?
BORED/PP+GET/PPST-2SG
Got-you_bored?
You got bored?

- *نه [خسته] نشدم.*
Na khaste nashodam.
NEG BORED/PP+NEG+GET/PPST-1SG
No, not-got-I_bored.
No, I didn´t get bored.

As noted in the above three examples, all the English sentences contain verbal ellipsis; however, the occurrence of VPE is only possible in Persian for example 0 and example (6), because, in the former, the modal verb *میتونم* (mitoonam) [CAN/MOD/PRS-1SG] 'can-I' ('I can') can take the function of the verb phrase *زندگی تنها کنم* (tanha zendegi konam) [ALONE LIFE/NC_DO/LV/PRS-1SG] 'do-I_live-I alone' ('live alone') and lead to ellipsis, and in the later the passive operator *شدن* (shodan) 'get' can take the function of the verb.

In Persian light verb constructions, the ellipsis of the nominal component (NCE) may occur. In example (7) below, the antecedent verbal group *تامین کردن* (Taamin kardan) [SUPPORT/NC+DO/LV] 'do_support' ('to support') is a light verb construction consisting of the nominal component *تامین* (taamin) 'support' and the light verb *کردن* (kardan) 'to do' Here, omitting the nominal component and keeping the light verb leads into NCE.

(7) *I am supporting you all.*
- *Well, don't* [*support us*].

HT
دارم همتونو تامین میکنم.
Daram hamatoono taaminl mikonam.
HAVE/PRG-1SG ALL+YOU/ACC-2PL SUPPORT/NC+DO/LV/PRS-1SG
Have-I all you do-I_ support.
I am supporting you all.
خوب [تامین] نکن.
Khob [taamin] nakon.
WELL[SUPPORT/NC+] +NEG+DO/LV/IMP-2SG
Well not- do [*support*].
Well don't do [*support*].

VPE after auxiliaries *do*, *be*, and *have* (2,559 cases) are translated by human translators as follows:

i. If the English verbal group is a lexical verb in Persian (which happened in 22.24% of the cases), the gap produced by VPE is filled by its antecedent (13.46%), as in example (4); or b) by a pro-verb (8.78%), as in example (8).

(8) *We're not leaving you.*
- *Yes, you are* [*leaving me*].

HT
ما تنهات نمیزاریم.
Ma tanhat nemizarim.
WE/NOM ALONE+YOU/ACC NEG+LEAVE/PRS-1PL
We not leave-we alone-you.
We don't leave you alone.

دارین همینکارو میکنین.
Darin haminkaro mikonin.
HAVE/PRSPR-2PL+THIS/DET+WORK/ACC+DO/PRS-2PL
You have-you do this work.
You are doing so.

ii. If the English verbal group is a light verb construction in Persian (45.37%), besides the above two forms of translation (the recovery of the antecedent verb (9.66%) and replacing a pro-verb (8.49%)), there is a third form that contains NCE) (27.22%) (example (7)).

iii. For VPE passive voice after auxiliary *be* the VPE can be retained (1.13%), as in example (6).

iv. In case of confirming a previous statement, the translation can be a confirming adverb (23.02%), as همینطوره (hamintore) [LIKEWIES/ADV+BE/PRS-3SG] 'likewise is' ('it is likewise') (example (9)), البته (albate) 'of course', حتما (hatman) 'certainly', etc. with or without بله (bale) 'yes' preceding them.

(9) *I thought you hated Kelso.*
- *I do.*

HT
فکر میکردم از کلسو متنفري.
Fekr mikardam az kelso motenaferi.
THOUGHT/NC+DO/LV/PST-1SG FROM KELSO HATE/PRS-2SG
Did-I_thought hate-you from Kelso.
I thought you hate Kelso.

- همینطوره.
Hamintore.
LIKEWISE/ADV+BE-3SG
Is likewise.
It is likewise.

v. With English VPE in comparative and adverbial clauses, the clauses are reduced into Persian adverbial or comparative phrases (4.30% (examples (10) and (11), respectively); that is, the VPE is kept.

(10) *He gives you more courage than I do.*

HT
او بیش از من به تو جرات میده.
Oo bish az man be to jorat mide.
HE/NOM MORE THAN I/NOM TO/PRE YOU/NOM-2SG COURAGE/NC+GIVE/LV/PRS-3SG
He gives_courage to you more than I.
He gives you more courage than I do.

(11) *We should not move before they do*.

HT

ما نباید قبل از اونا هیچ اقدامی کنیم.

Ma nabayad ghabl az oona hich eghdami konim.

WE/NOM NEG+MUST/MOD BEFORE-OF THEM/NOM NOTHING MOVEMENT/ NC+ DO/LV/ PRS-1PL

We not-must do-we_movement nothing before of them.

We must not do anything before they do.

vi. The sentences with VPE can be translated non-literally. In some cases the translators did not feel obliged to keep the same structure as the original; while keeping the content, they did not keep the original wording. This tendency may be justified for the purpose of making the text more natural and appropriate, especially in the conversational (oral) situation of subtitles, and for creating diversity and expressiveness, or even for conveying the ironic sense of the sentence (example (12)) (These cases are called, hereafter, 'non-literal'). This approach, however, is only implemented in a few instances (2.38%). Therefore, it can be claimed that the patterns studied here were, for the most part, quite close to those of the original structures.

(12) *I find him*.

- *You do*?

HT

من پیداش میکنم.

Man peydash mikoanm.

I/NOM FINDING/NC+HIM/ACC+DO/LV/PRS-1SG.

I do-I_finding-him.

I find him.

- هیچ کس هم نه تو.

Hichkas ham na to.

NOBODY TOO/PRO NO YOU/NOM-2SG

No nobody too you.

Nobody else, just you.

In the above example, if *You do?* had been translated literally, keeping the ellipsis, as in تو؟ (to) 'you?', or by replacing a pro-verb, as تو اینکارو میکردی؟ (To in karo mikardi?) [YOU-2SG THIS/DET WORK/NC+DO/PRS-2SG] '*You do-you this work?*' (*You do that?*), the ironic sense of the answer would have been distorted.

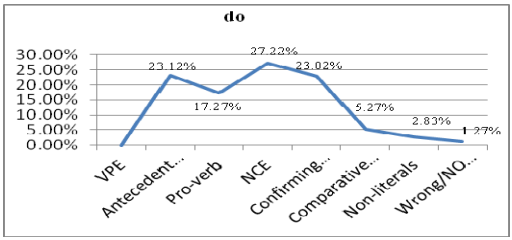
Tables 3a and 3b (Table 3a continues in Table 3b), and Figure 1, below, sum up these observations.

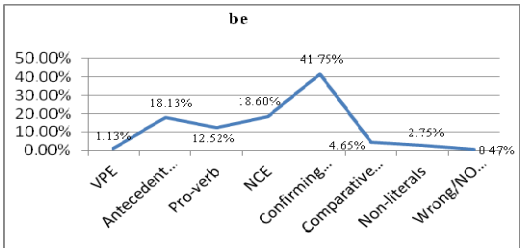
Table 3a: HT of VPE after operators (*do*, *be*, *have*)

Operator	Cases	Lexical verbs			Light verb constructions			
		VPE %	Antecedent %	Pro-verb %	VPE %	Antecedent %	Pro-verb %	NCE %
do	1,025	–	13.46	8.78	–	9.66	8.49	27.22
be	1,054	0.85	12.15	8.54	0.28	5.98	3.98	18.60
have	480	–	36.87	2.92	–	16.88	5.43	19.18
Total	2,559	0.35	17.31	7.58	0.12	9.50	6.06	22.16

Table 3b: HT of VPE after operators (*do*, *be*, *have*)

Lexical verbs/light verb constructions				
Operator	Confirming statements %	Comparatives/Adverbials%	Non-literals %	Wrong/No Translation %
do	23.02	5.27	2.83	1.27
be	41.75	4.65	2.75	0.47
have	16.63	1.47	0.62	–
Total	29.54	4.30	2.38	0.70





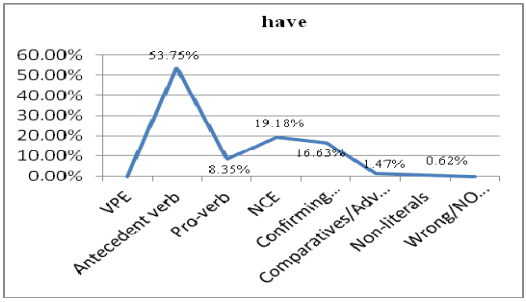


Figure 1: HT of VPE after operators ('do, be, have')

As Table 3 (a and b) and Figure 1 show, the HT tendency in translation of VPE after the operators is mostly toward keeping the original structure (Table 3a: 63.08%) rather than following the other structures (Table 3b: 36.92%). VPE after auxiliary *have* is mostly recovered in Persian by the antecedent verb (53.75). The majority instances of VPE after auxiliary *be* occured in a confirming answer to a previous statement; for which the adopted translation strategy was using the confirming statement (41.75%). The strategies for other instances of VPE after this auxiliary were quite similarly divided among the strategies of the antecedent verb (18.13%), pro-verb (12.52%), and NCE (18.60%). It is worth mentioning that NCE is, in fact, a sub-category of VPE, as the light verb component is deleted and the nominal component is retained. VPE after auxiliary *do* was translated using approximately the same number of occurrences of strategies: antecedent verb (23.12%), pro-verb (17.27%), NCE (27.22%), and confirming statement (23.02%). A residual number of cases were translated using different constructions (comparative/ adverbials) or were translated non-literally.

Regarding GT, it does not recover the gap resulting from VPEs. It also translates all the operators *do*, *be*, and *have*, as a lexical or light verb; it lacks person and number agreement between the subject and the verb; and the tense is not preserved, as in example (13):

(13) *I love you, I always have* [*loved you*].

GT

*من شما را دوست دارم ، من همیشه داشته باشد.**

Man shoma ra doost daram, man hamishe dashte bashad.

I/NOM YOU/ACC(H) LOVE/NC_HAVE/LV/PRS-1SG I/NOM ALWAYS HAVE/PP+B/INF-3SG

I have-I_love you, I always had-be.

I love you, I always has.

Persian allows VPE for the modal verbs *can*, *may*, and *must*/*have to*, if they are translated as مجبور بودن (majboor boodan) [OBLIGED+BE/GR]). Accordingly, all the possible translation strategies mentioned above for VPEs after operators are also possible for VPEs after these modal verbs. However, English VPEs after the modal verbs *will*, *should*, and *must*/*have*, if they are translated as باید (bayad) 'must', cannot be translated into Persian by keeping the ellipsis.

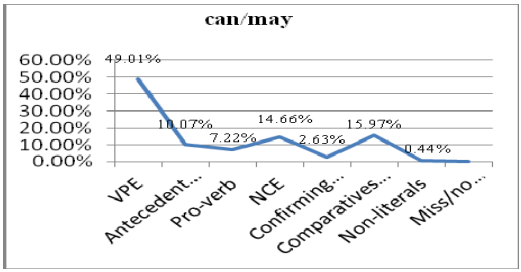
Tables 4a and 4b, and Figure 2, below, sum up the translation strategies for modal verbs:

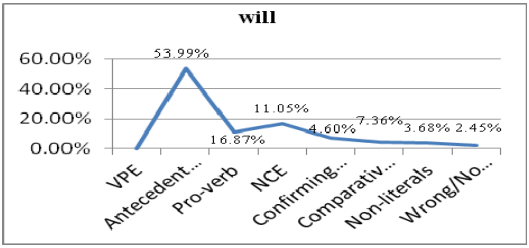
Table 4a: HT of VPE after modal verbs

Modals	Cases	Lexical verbs			Light verb constructions			
		VPE %	Antecedent %	Pro-verb %	VPE %	Antecedent %	Pro-verb %	NCE %
can/may	457	28	7.44	2.84	21.01	2.63	4.38	14.66
will	326	–	35.28	4.30	–	18.71	6.75	16.87
must/have to/should	352	10.51	19.03	9.66	17.05	17.90	4.55	9.09
Total	1,135	14.54	19.03	5.38	13.75	11.98	5.11	13.57

Table 4b: HT of VPE after modal verbs

Lexical verbs/light verb constructions				
Modals	Confirming statements %	Comparatives/Adverbials %	Non-literals %	Wrong/No Translation %
can/may	2.63	15.97	0.44	–
will	7.36	4.60	3.68	2.45
must/have to/should	10.22	1.99	–	–
Total	6.34	8.37	1.23	0.70





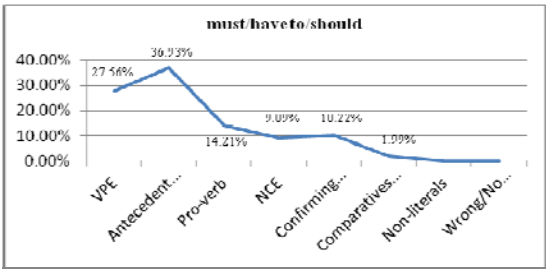


Figure 2: HT of VPE after modal verbs

According to Table 4 (a and b) and Figure 2, here again, the HT tendency in translation of VPE after the operators, is more toward keeping the original structure (Table 4a: 83.36%)) rather than following the other structures (Table 4b: 16.64%). The structures with *can*/*may* were mostly translated into Persian by keeping the VPE (49.02%); the structures with *will* were mostly translated using the antecedent verb (50.99%); and the structures with *must*/*have to*/*should* kept the VPE for 27.56 per cent of cases and used the antecedent verb in translation for 36.93 per cent of cases.

GT in dealing with VPE after modal verb *can* operates fairly well (example (14)); however it fails in dealing with other modal verbs: after *will* the translation is 'subject + modal verb the passive voice'; after after *may* and *must*/*should* it gives the unnatural combination of subject and the modal verbs following it, as the gap needs to be recovered by the antecedent verb or replaced by a pro-verb.

(14) *Laugh while you can* [*laugh*].

GT

بخند در حالیکه شما می توانید [بخندید].

Bekhand dar halike shoma mitavanid [bekhandid].

LAUGH/IMP WHILE YOU/NOM-2SG(H) CAN/MOD/ PRS-2SG(H) [LAUGH/INF-2SG(H)]

Laugh while you can-you [*laugh*].

Laugh while you can [*laugh*].

i. *VPE after complementizer 'to'*

English VPEs occurring after infinitival complementizer *to* (654 cases) are mostly translated by filling the gaps with the full verb (513 cases), as in example (15); in 127 cases, the ellipsis was kept; and in a residual number of cases (12 instances) the gap was replaced by a pro-verb.

(15) *I kept my mouth shut because they wanted to*. (That is ...*they wanted* [*me*] *to* [*keep my mouth shut*]; or …*they wanted* [*me*] *to* [*do so*].)

HT

اونا ازم خواستن که حرفی نزنم.

Oona azam khastan ke harfi nazanam.

THEY/NOM FROM/PRE+ME WANT/PST-3PL THAT/CNJ WORD/NC+ANY/DET+NEG+ HIT/LV/INF-1SG

They wanted-they from me that not-hit_any-word.

They wanted me that I do not talk.

GT retains this kind of ellipsis in all cases; hence, the output is often unnatural even if interpretable in most cases. Consider the Google translation of the example (15).

GT

من دهانم را بسته نگه داشته چراکه آنها می خواستند [دهانم را بسته نگه دارم].

Man dahanam ra baste negah dashte chera ke anha mikhastan [dahanam ra baste negah daram].

I/NOM MOUTH+MY/ACC SHUT/PP KEEP/NC+HAVE/LV/PPER BECAUSE THEY WANT/PST-3PL [MOUTH+MY/ACC SHUT/PP KEPT/NC +HAVE/LV/INF-1SG].

I have-had_kept my mouth shut because they wanted-they [*have-I_ kept shut*].

I kept my mouth shut because they wanted [*I keep my mouthshut*].

The output would be more natural if, like the HT of the same example, the gap after the verb می خواستند (*mikhastand*) had been filled with the antecedent verb or pro-verb, as is the most common process in Persian.

ii. VPE with pro-form constructions

English VPEs with pro-form structures with *so*/*too*/*as well*/*neither*/*either* (175 cases) were translated into Persian, by following Persian pro-form structures; thus, like English, the ellipsis was kept. Only in a residual number of cases the pro-form was followed by the antecedent verb or by a pro-verb, as in (example (16)) using the antecedent verb مردن (*mordan*) 'die':

(16) *When she died* so *did I.*

HT
وقتي اون مرد من هم مردم.
Vaghti oon mord mna ham mordam.
WHEN/RPRO SHE DIE/PST-3SG I/NOM ALSO DIE/PST-1SG
When she died I too died-I.
When she died I died too.

It seems that GT, in dealing with this particular VPE, mostly produces inadequate translations. From among 67 cases of VPE in pro-form structures with *so*, only 12 cases were properly translated, using an adequate pro-form structure; and from 20 VPE instances in pro-forms with *too*, 14 cases were translated by using equivalent pro-forms. In translating the VPEs in other pro-forms, GT produced noise. From the collected evidence, it was not possible to discover why GT only performs properly in some cases.

5 Assessment

The same corpus, which was used for the data analysis, was used in order to verify the representativeness of the data presented above against the corpus. The task was carried out on six random locations of the corpus (each containing 5% of the corpus, in total 183,607 sentences). In total, 1,094 cases of VPE instances were weighed up. To sum up, not much difference was observed between the results of the data analysis and those of the assessment. Therefore, we can claim that, as far as this corpus is considered, the results are relatively stable. The details of the assessment procedure and the obtained results have been presented in Shahabi (2011).

6 Conclusion

The results indicate that the Persian human translator dealing with English VPE predominantly adopts the strategy of recovering the zeroed verb from its previous occurrence in the discourse. Naturally, in some cases, instead of a verb, a pro-verb is used. For light verb constructions in Persian, the light verb is retained and the nominal component is zeroed. For a residual number of cases the strategies were non-literal.

This general behavior, however, depends on the auxiliary verb used in the text. Differences in the auxiliary verb used in English VPE have a relevant bearing on the choice of the strategies the human translator adopts. For instance, the translation strategy for VPE after modal verbs *can*, *may*, and *must*/*have to*, VPE in pro-form structures is mostly to keep the ellipsis in the text, while for VPE after other modal verbs and after operators, the strategy is to fill the gap with the antecedent verb or a pro-verb.

As a statistical-based MT system, GT does not take into consideration the discourse previous to the sentence under processing. The system, therefore, is incapable of recovering the gap induced by English VPE, and this results in incorrect translation output, as has also been confirmed by the analysis.

The comparison between HT and GT of Persian texts indicates that a stronger effort should be invested in an anaphora resolution module, particularly for certain English VPE patterns: those involving auxiliary verbs *do*, *be*, *have*, and *will*, and those after complementizer *to*.

7 Notes

Morphological analysis appearing in abbreviation form are listed as follows: ACC: Accusative; ADV: Adverb; CNJ: Conjunction; DET: Determiner; H: Honorific; IMP: Imperative; INF: Infinitive; INT: Interogative; LOC: Locative; LV: Light verb; MOD: Modal; NC: Nominal component; NEG: Negation; NOM: Nominative; PL: Plural; PRO: Pro-form; PP: Past Participle; PPRS: Passive & Present; PR: Progressive; PRE: Preposition; PRS: Present; PRSPR: Present Progressive; PST: Past; RPRON: Relative Pronoun.

8 Acknowledgement

The research for this paper was supported by Erasmus Mundus Masters in NLP & HLT programme.

9 References

De Vries, Dennis. 2009. A semantic approach to antecedent selection in VP ellipsis. M.A. thesis. Nederlands: University of Groningen.

Halliday M.A.K. and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Hardt, Daniel. 1997. An empirical approach to VP ellipsis. *Computational Linguistics* 23(4): 525–541.

Lappin, Shalom. 1996. The interpretation of ellipsis. In S. Lappin (ed.). *Handbook of contemporary semantic theory*, 145–175. Oxford: Blackwell.

Lappin, Shalom. 2005. A sequenced model of anaphora and ellipsis resolution. In A. Branco, A. McEnery and R. Mitkov (eds*.). Anaphora processing: Linguistic, cognitive, and computational modeling*, 3–16. Amsterdam : John Benjamins.

Lappin, Shalom and Benmamoun Elabbas (eds.). 1999. *Fragments: Studies in ellipsis and gapping*. New York: Oxford University Press.

Lappin, Shalom and Michael McCord. 1990. Anaphorical resolution in slot grammar. *Computational Linguistics* 16(4) : 197–212.

Lobeck, Anne. 1995. *Ellipsis*. Cambridge: Cambridge University Press.

Malmkjaer, Kirsten. 1998. Love thy neighbor: Will parallel corpora endear linguists to translators? *Meta* 43(4): 534–541.

Mitkov, Ruslan. 2002. *Anaphora resolution*. London: Longman.

Mitkov, Ruslan, Branimir Boguraev and Shalom Lappin (eds.). 2001. Introduction to special issue on computational anaphora resolution. *Computational Linguistics* 27(4) : 473–477.

Mitkov, Ruslan, Evans Richard and Orasan Constantin. 2002. *A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method.* Proceedings of CICLing-2000, Mexico City, Mexico, 168–173.

Nielsen, Leif Arda. 2005. A corpus-based study of verb phrase ellipsis. Unpublished Ph.D. thesis. London: University of London.

Paumier, Sébastien. 2008. *Unitex 2.0 user manual*. Paris: University of Marne-la-Vallée (UMLV).

Pilevar, Mohammad Taher. 2010. *Tehran English Persian parallel corpus*. http://ece.ut.ac.ir/nlp/resources.htm (Accessed 4 April 2010).

Python Programming Language- Official Website.http://www.python.org/ (Accessed 20 October 2010).

Shahabi, Mitra. 2011. A corpus-based translation study on English-Persian verb phrase ellipsis. Unpublished M.A. thesis. Faro (Portugal)/Wolverhampton (United Kingdom): University of Algarve/University of Wolverhampton.

Toury, Gideon. 1994. *Descriptive translation studies and beyond*. Amsterdam: John Benjamins.

Unitex. http://www-igm.univ-mlv.fr/~unitex/ (Accessed 4 October 2010).