International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:11, No:2, 2017

Study of Syntactic Errors for Deep Parsing at Machine Translation

Yukiko Sasaki Alam, Shahid Alam

***Abstract*—**Syntactic parsing is vital for semantic treatment by many applications related to natural language processing (NLP), because form and content coincide in many cases. However, it has not yet reached the levels of reliable performance. By manually examining and analyzing individual machine translation output errors that involve syntax as well as semantics, this study attempts to discover what is required for improving syntactic and semantic parsing**.**

***Keywords*—**Machine translation, error analysis, syntactic errors, knowledge required for parsing.

I. Introduction

ROBUST syntactic parsing improves the quality of many NLP applications such as machine translation (MT), information extraction, text summarization, and question answering. As has long been observed, form and content go hand in hand [5], [8]-[10], [14]. The improvement of syntactic parsing enhances the quality of understanding sentences and texts.

Although the quality of syntactic parsing has been improving, it has not achieved reliable performance yet. Reference [20] reports that the current accuracy of parsing, whether constituency or dependency, is in the range of 80-84%. The results of our study, however, indicate much lower accuracy. More than two-thirds of 226 output sentences we examined contained parsing errors [29]. The average intelligibility of the output sentences was 1.19 out of 4, the maximum score. That means that on average, output sentences are "unintelligible in spite of several intelligible words and phrases." While this low intelligibility is partially due to lexical errors, it is also due to parsing errors. In fact, both types of errors often influence each other, making it hard to tell which caused which.

This paper aims to discover what knowledge is required for improving syntactic parsing by manually examining MT output errors that mainly involve syntactic parsing.

II. Related Study

Past research has placed emphasis on the categorization of MT output errors rather than focusing on and examining the causes of individual syntactic errors. Error categories recognized in the past are such as elision, word order, conjunction, and clause boundary. A detailed discussion of each category of errors and the probable causes has not been made, however, because it was not the intention of the past research. Reference [7] presented a variety of error categories, and grouped them into three levels respectively according to improvability and intelligibility [7], but no discussion was made on the causes. Automatic error analysis has also brought general tendencies of errors to light, but not shown what is really happening to individual errors and what are plausible remedies for these errors [3], [6], [25], [26].

This study investigated each output, identifying syntactic errors, and attempting to find out what problems confront syntactic parsing.

Yukiko Sasaki Alam, Ph.D. is a full professor at the department of digital media at Hosei University, Tokyo, Japan (e-mail: sasaki@hosei.ac.jp).

Shahid Alam is an independent researcher. (e-mail: shahid.alam@ gmail.com).

III. Data

The collected data comprises 226 output sentences translated by an online English-Japanese translation [28]. The input sentences are taken from seven articles on the stock market in online news and financial magazines posted in early 2016. The average length of the input sentences is 19.6 words, [1] with the shortest of four words and the longest of 48 words. Each output sentence was examined manually in order to find syntactic, morphological and lexical errors. This paper deals with errors related to syntactic parsing.

IV. Results

This section shows the results of our study: knowledge required for avoiding errors involving syntactic parsing. It should be noted, however, that some errors were not caused by the absence of one area of knowledge only, but of more than one. For instance, a failure to recognize a relative clause may have been due to a failture to recognize a clause boundary or the argument structure of the predicate of the clause, or both.

A. Idiomatic Constructions

English, like other languages, allows for a large number of idiomatic sentence and phrase constructions. Without knowledge about these idiosyncratic constructions, a MT system is unable to parse sentences and phrases containing them, because they are not as compositional as regular sentences and phrases, and thus grammatical rules cannot handle them. Input sentences in our study contained many such constructions. They can be divided into two groups, those consisting of components without any intervening elements between them, and those of components with other elements between them. The former is called *continuous idiomatic constructions* while the latter *discontinuous idiomatic constructions*. The following will discuss both types that the MT system in question failed to recognize.

[1] A multiword such as *take care of* is counted as three-word long.

1) Continuous Idiomatic Constructions

Table IA illustrates example sentences made of continuous idiomatic constructions that the system failed to identify. The first example numbered (1) of Table IA is:

… is for two stocks to rise to analysts' average price targets

The *for*-NP-*to*-VP construction is a clause semantically consisting of the subject noun phrase (NP) *two stocks* and the predicate verb phrase (VP) *rise to analysts' average price targets*, and the failure to identify each functional role of the NP and the VP respectively as subject and predicate led to an unintelligible output sentence.

An example in (2) contains a NP which is comprised of *all*, followed by a relative clause beginning with the subject NP *the index*, followed by the predicate verb *needs*, followed by the purpose clause *in order to cross …*:

And at this point, all the index needs in order to cross that elusive level is …

In (2), *all* is the antecedent of the relative clause *the index needs in order to cross that elusive level*. It is also the object NP of the predicate *needs*. The bracketing of the NP beginning with *all* and ending with *level* is:

NP [*all* + RELATIVE CLAUSE [*the index* + *needs* + missing object NP + ADVERBIAL CLAUSE [*in order to cross that elusive level*]]].

The MT system failed to recognize this idiomatic relative clause construction with *all* its antecedent, and incorrectly grouped together *all the index* as a NP, causing one error after another.

A similar example involving a relative clause with a special antecedent is (3):

What Honeywell saw was the chance to wring from United …

In (3), *what* is the antecedent of the relative clause *Honeywell saw*, as well as the object of the predicate verb *saw*. The bracketing of the NP beginning with *What* is:

NP [*What* + RELATIVE CLAUSE [*Honeywell* + *saw* + missing object NP]].

The system failed to recognize the NP consisting of the antecedent NP *what*, followed by the relative clause *Honeywell saw*. As a result, the predicate verb *saw* was mistakenly identified as a noun that stands for a hand tool for cutting wood. Due to the failure to parse the initial NP of the sentence, the remaining components of the sentence were not assigned the correct functional roles, resulting in a gibberish output sentence.

An example in (4) is a NP consisting of the head NP *low risk*, followed by the appositive *that*-clause:

low risk that the dividend could be cut

The bracketing is:

NP [NP [*low risk*] APPOSITIVE CLAUSE [*that* NP *could* VP]].

The appositive *that*-clause is a full clause without any missing argument of the predicate verb (*be cut*). In order to parse such a NP consisting of the head NP followed by the appositive clause, the MT system should know if it is a clause with a missing argument (like a relative clause) or a full clause (like an appositive *that*-clause). As the head NP in this construction refers to an abstract entity such as *low risk*, *fears*, *claim*, *fact*, and *expectations*, the use of a simple semantic feature system will be helpful: whether it refers to an abstract entity or not

A similar construction to (4) is (5), but the difference from (4) is without the complementizer *that* for the appositive clause:

one reason J&J is one of just three U.S. firms with …

The bracketing of this NP is:

NP [NP [*one reason*] APPOSITIVE CLAUSE [NP VP]].

This type of appositive clause can be without a complementizer such as *why* and *that*. The omission of the complementizer is limited to such head nouns as *reason*, *way*, and a noun of time. Therefore, the system should look for such a noun when it finds an appositive clause without a complementizer.

TABLE IA
CONTINUOUS IDIOMATIC CONSTRUCTIONS THE SYSTEM FAILED TO RECOGNIZE

(1) *for* + NP+ *to* + VP	*is for two stocks to rise to analysts' average price targets*
(2) *all* + clause having a missing argument[2]	*And at this point, all the index needs in order to cross that elusive level is …*
(3) *what* + clause with a gap	*What Honeywell saw was the chance to wring from United*
(4) NP [abstract] + *that* + clause	*low risk that the dividend could be cut*
(5)… *reason* + clause	*That immense cash-generating power is one reason J&J is one of just three U.S. firms with the top, triple-A credit rating*
(6) *With* NP + VP*ing*[3]/VP*en*[4], clause	(a) *With the stock yielding just 1.7%, CVS's appeal is long-term income growth, not current yield.* (b) *But with the shares richly priced, at 20 times estimated 2016 earnings per share, investors need to have a long-term view.*
(7) *given* + NP + PP, clause	*Given the $216 billion in cash and investments in Apple's treasury, if any dividend is safe it's this one.*
(8) *that's where* X *could come in*	*And that's where Apple could come in.*
(9) *how* + adjective/adverb	*…, you look at how desperately they worked to keep the stock market up.*
(10) *no other* + noun	*No other country has displayed that.*
(11) *so* + adjective/adverb phrase + *that* + clause	*… so well in 2016 that many are starting to look pricey*
(12) *too* + ADJ/ADV[5] phrase + *to* + VP	*China is far too volatile and murky to invest in.*

[2] *Clause having a missing argument* means that an argument of the predicate verb such as the subject NP and the object NP is missing.
[3] *VPing* denotes a verb phrase (VP) beginning with the present participle of a verb or a gerund.
[4] *VPen* stands for a VP beginning with the past participle of a verb.
[5] *ADJ* stands for adjective while *ADV* for adverb.

TABLE IB
DISCONTINUOUS IDIOMATIC CONSTRUCTIONS THE SYSTEM FAILED TO RECOGNIZE

(1) *the* + comparative form of ADJ …, *the* + comparative form of ADJ …	*For this reason, the higher the share price of a given stock, the more importance it has in the index.*
(2) *not just …, but*	*Brokerage Goldman Sachs goes a step further, suggesting that Exxon won't just hold the payout steady but could well boost it*
(3) Imperative clause, *and* + clause	*Couple that with heated competition from rivals such as…, heavy capital spending on new technology and a high debt load, and it isn't surprising that Verizon warns that it's hitting a profit wall.*

(6a) in Table IA illustrates a circumstantial *with*-prepositional phrase (*with*-PP):

With the stock yielding just 1.7%, CVS's appeal is long-term income growth, not current yield.

The functional structure of the circumstantial *with*-PP is:

PP [*with* + subject NP + predicate VP], MATRIX CLAUSE [NP VP]

In (6a), the subject NP of the circumstantial *with*-PP is *the stock*, and the predicate VP is *yielding just 1.7%*. That is, a circumstantial construction describes a clause. When an output language such as Japanese does not have such a circumstantial phrase representing a clause, the grammatical functions of subject and predicate must be explicitly indicated in the output sentence. The MT system failed to recognize these grammatical functions expressed in this circumstantial construction, resulting in an ungrammatical output sentence.

(6b) is another circumstantial *with*-PP in a sentence, but it differs from (6a) in that the predicate VP begins with the past participle of the verb (VP*en*) instead of VP*ing*. To parse (6b) is more difficult than (6a), because the verb, *priced*, is pre-modified by an adverb, *richly*, as well as post-modified by a long adverbial *at*-PP, *at 20 times estimated 2016 earnings per share*.

An example in (7) shows a sentence-initial PP in the form of [*Given* + NP], which means 'taking NP into account':

Given the $216 billion in cash and investments in Apple's treasury, if any dividend is safe it's this one.

The word *given* in this construction can be treated as a preposition. The bracketing of this PP is PP [P + NP].

2) Discontinuous Idiomatic Constructions

It is less difficult to parse continuous idiomatic constructions than discontinuous idiomatic constructions. Continuous idiomatic constructions consist of a predetermined order of constituents, and allow for much less intervention between the constituents [2], [17], [19]. However, in order to recognize discontinuous idiomatic constructions, the parser sometimes has to cross over clause boundaries when looking for the member constituent. The components of discontinuous idiomatic constructions can be far apart, often intervened by other words. This means that the number of constituents of a sentence to identify is much larger than for continuous idiomatic constructions.

In (1) of Table IB, the sentence construction involves the following phrase and clause boundaries:

SENTENCE [CLAUSE1[NP[*the higher*] NP[*the share price of a given stock*]], CLAUSE2[NP[*the more importance*] NP[it] VP[V[*has*] PP[*in the index*]]]

As the bracketing shows, the sentence structure in (1) is complex with a missing BE verb in the first clause. Inversions are found in each clause: the complement (= *higher*) appears before the subject NP (= *the share price of a given stock*) in the first clause, while in the second clause, the object NP (= *the more importance*) is moved to the clause-initial position. Looking at this example alone, it is easy to understand the difficulty in parsing a discontinuous idiomatic construction.

An example in (2) of Table IB illustrates a discontinuous idiomatic construction *won't just … but …* The bracketing of this construction is:

SENTENCE [NP [*Exxon*] VP1 [*won't just hold the payout steady*] CON [*but*] VP2 [*could well boost it*]]

The conjunction *but* conjoins the first and second VPs. It should not be treated as the normal usage for *but*, because the meaning of the second VP is semantically not in contrast to that of the first VP. The MT system failed to identify this idiomatic construction, and translated *but* into its regular meaning, resulting in a contradictory output sentence.

In (3) of Table IB, the conjunction *and* is a part of the idiomatic construction starting with an imperative clause, followed by a comma, followed by the conjunction *and*, followed by a declarative sentence, as in *Hurry up, and you will be in time for the meeting.* Therefore, it should not have been translated literally as the regular *and*, but it was, resulting in an awkward translation.

As understood from the discussion of a few examples of discontinuous idiomatic constructions, success in the identification of them also depends on the parser's basic skills: to parse sentences into grammatical phrases and clauses. Without these basic abilities, it will not be possible to recognize idiomatic constructions, whether continuous or discontinuous.

B. Domain-Specific Constructions

Table II illustrates a few input sentences comprised of domain-specific constructions that appeared frequently in the texts on the stock market. They describe an entity's change of quantity over time (like (1) and (3) in the table) as well as the comparison of two or more corporations in terms of the quantities of the related entities (like (2)). In particular, the sentence construction in (1) is a typical one in a text on the stock market, showing an increase, but the system failed to understand the simple construction. It will be a good idea to collect sentences or verb phrases representing domain-specific events so that they can be handled in a more unambiguous manner. The meanings of words are often domain-specific and easy for a machine to predict.

TABLE II
DOMAIN-SPECIFIC CONSTRUCTIONS THE SYSTEM FAILED TO RECOGNIZE

(1) *Hain was up 1.7 percent in 2016.*
(2) *Adjusted earnings before interest, depreciation and amortization for the fiscal fourth quarter, which ended March 31, was $2.16 billion, compared with a $2.02 billion average of estimates compiled by Bloomberg.*
(3) *The company cut $1.3 billion in expenses in the fiscal year.*

C. Multiword Expressions (MWEs)

Ever since the publication of [23], more and more research has been devoted to MWE problems that challenge NLP [1], [4], [13]-[16], [18], [20], [21], [24]. A MWE refers to a compound of words which should not be translated compositionally, but be treated as a whole representing one meaning.

Knowledge on nominal MWEs is vital not only for understanding the correct meanings, but also for syntactic parsing. For instance, the system failed to identify *track* as a half part of *track record* in the phrase *its operational efficiency, track record of execution and rational capital allocation decisions*, and treated it as a verb, resulting in an incorrect output denoting that its operational efficiency tracks the records of execution and rational capital allocation decisions. Many similar errors were detected. The system's failure to recognize MWEs led to the incorrect identification of the parts of speech of the words, causing parsing failures. The MWEs listed in Table III are respectively a phrasal verb *bring in* in (1), an adjective phrase *as low as* in (2), a proper noun *Doug Kass* in (3), a nominal compound *track record* in (4), and an adverbial phrase *all the way back* in (5). It will improve the accuracy of parsing to prepare a list of MWEs that appear frequently in texts in a particular domain.

TABLE III
MWES THE SYSTEM FAILED TO RECOGNIZE

(1) phrasal verbs	*That product portfolio brought in $30.3 billion in sales in 2015.* (not as PP, *in $30.3 billion*)
(2) *as* + ADJ + *as* + NP	*The dividend provided a firewall, of sorts, as the Wal-Mart shares crumbled to as low as $56.30 in December.* (*as* not in the capacity of …)
(3) proper names	*However, the specter of impending competition was enough for Real Money Pro's Doug Kass,…*
(4) nominal compounds	*Its operational efficiency, track record of execution and rational capital allocation decisions* *It also narrowed its annual sales forecast to $2.95 billion to $2.97 billion, compared with a previous projection of $2.9 billion to $3.04 billion.*
(5) adverbial compounds	*So I went back and ran more numbers on the "death cross" going all the way back to the 1920s,*

D. Clause Boundaries

As the MT system in question is a statistical MT (SMT), and no constituency parser seems to support the system, it performed poorly in segmenting sentences into grammatical units such as phrases and clauses. The inability to recognize phrase or clause boundaries resulted in the failure to deal with complex sentences consisting of multiple clauses.

A sentence construction, which comprises a matrix clause, followed by a comma, followed by a VP*ing* as in (1) of Table IV, frequented the texts we examined, and the system failed to parse almost all of those. Most sentences consisting of a matrix clause followed by a VP*ing* clause are long, since they consist of more than one clause. Three such examples are listed in (1a-1c) of the table. Most of the present participles of verbs (V*ing*) in this construction were mistaken for nouns, resulting in serious parsing errors. The comma immediately preceding a VP*ing* plays a significant role of indicating the end of a matrix clause, but the system did not recognize its role. This failure suggests that it was not only unaware of this particular construction, but also unable to identify a matrix clause.

TABLE IV
INPUT SENTENCES INCORRECTLY TRANSLATED DUE TO THE FAILURE TO IDENTIFY BOUNDARIES

(1) clause, VP*ing* or clause VP*ing*	(a) *The Dow is a price-weighted index, meaning it replicates …* (*Meaning* is not a noun.) (b) [matrix clause], *topping the $734.6 million analysts had predicted* (*Topping* is not a noun) (c) *China's stock market has dropped from a June 12 peak wiping out almost $4 trillion in value in less than a month …*
(2) clause, VP*en*	*Earnings hit a record in 2014 before slipping in 2015, hurt by weakness in some of the device businesses and by the strong dollar.*
(3) Relative clause (= antecedent NP followed by a clause with a missing argument/element)	(a) *the cash its businesses generate after subtracting capital expenditures* (b) *The foodmaker posted $749.9 million in fiscal third-quarter sales on Wednesday, topping the $734.6 million analysts had predicted* (c) *57 percent of the analysts who cover the stock rate it as a "buy,"*
(4) *If* + clause, *as* + clause, clause	*If the dollar continues to weaken against many foreign currencies, as it has so far this year, 3M's bottom line should benefit.*
(5) *I think* + clause, *and* + *that* + clause	*I think AAPL's future sales-and-profit outlook is worse than consensus expectations, and that the tech giant's valuation faces numerous headwinds.*
(6) failure to parse the dependent clause	(a) *… after investors who borrowed to buy shares had to unwind trades.* (b) *the buffer provided by the company's dividend probably helped keep the stock from a deeper decline as growth fears escalated*

Although a comma usually preceded a VP*ing*, we found a sentence without a comma between the matrix clause and the following VP*ing* clause, as in (1c), which requires the system to recognize this multi-clause sentence without the help of a comma. More robust parsing is required to deal with this comma-less multi-clause sentence.

An example in (2) is similar to the previous construction (i.e. matrix clause + comma + VP*ing*), but differs in that the past participle of a verb (V*en*) occurs instead of a V*ing*. The treatment of this construction is the same in that the system needs to identify a matrix clause that precedes the VP*ing*.

Unlike the previous constructions with a dependent clause beginning with a V*ing* or V*en*, an example in (3) involves a relative clause. The success of identifying a relative clause depends on the system's ability to recognize a missing argument of the predicate of the relative clause. In (3), the relative clause is *its businesses generate after subtracting capital expenditures*, and it does not have the object NP of the transitive verb *generate*, but the system failed to recognize that, which caused the failure to identify its antecedent NP *the cash*.

The examples in (4), (5) and (6) also consist of multiple clauses, and the system was unable to identify the complex sentences, mainly because of the failure to recognize clause boundaries. An example in (5) reveals both the inability to

recognize clause boundaries and a lack of grammatical knowledge about the mandatory use of *that*-clause when the following clause is conjoined with *and* as the second object clause of a verb such as *think*. Without such grammatical knowledge, *that* in *that*-clause was wrongly identified as a demonstrative pronoun.

For a MT system to recognize clause boundaries, it should have knowledge about argument structures of predicates. Without this knowledge, it cannot understand whether a clause has a missing argument or not, as well as whether it is a full clause. The next section discusses examples directly involving knowledge about argument structures of predicates.

E. Argument Structures of Predicates

Most failures in the recognition of boundaries for phrases and clauses were due to the system' lack of knowledge about argument structures of predicates (namely, the arguments' participant roles in the events, such as the subject, object and Goal arguments as well as their basic semantic features). The system failed to recognize the argument structure of the verb *lift* in *lifted the annual rate to $2.08* in (1a) of Table V. The object NP *the annual rate* refers to an abstract entity, and the argument of the *to*-PP[6], *$2.08,* also refers to an abstract entity. In this argument structure, *lift* cannot mean 'raising a concrete object.' Knowledge about the argument structures of verbs (as well as verb patterns [12]) is required for proper syntactic parsing.

Deverbal nouns, when used to mean not products, but process, tend to inherit the same argument structures as the original verbs possess, and the system should be able to identify the argument structures of such deverbal nouns. For instance, in *a previous projection of $2.9 billion to $3.04 billion* in (2), the deverbal noun *projection* inherits the argument structure of the verb *project*, which governs such arguments as the subject NP, the object NP and the optional Goal argument (i.e. *to*-PP).

The object NP of *projection* is *$2.9 billion*, and the Goal argument is *to $3.04 billion*. As *projection* is morphologically a noun, and the *to*-PP (i.e. *to $3.04 billion*) modifies the deverbal noun, the *to*-PP is a nominal modifier. Unlike English, Japanese grammar demands a morphological distinction between modifiers of predicates (adverbial modifiers) and modifiers of nouns (nominal modifiers). Therefore, in Japanese, *to $3.04 billion* in the phrase *projection of $2.9 billion to $3.04 billion* should be marked as a nominal modifier by inserting a morpheme similar to the English *of*, as shown below in the Japanese word order using the English equivalents:

2.9 billion dollar-of 3.04 billion dollar-to-<u>of</u> projection

In order to add a morpheme to distinguish a nominal modifier from an adverbial modifier, the system needs to know if the modifier is nominal or adverbial. In most sentences containing such nominal modifiers, the MT system in question failed to distinguish them from adverbial modifiers, resulting in confusing output sentences. In addition, the system often failed to recognize the argument structure of a predicate that consists of the object NP, followed by its complement, as in *left it undervalued* in (4a) of Table V. The object NP's complement may take the form of the past participle of a verb as in (4a) or a PP as in (4b) and (4c), respectively *kept Exxon <u>in the black</u>* and *got you <u>out of the market</u>*. It could also be an adjective as in *keep the air <u>fresh</u>* or the present participle of a verb as in *keep us <u>going</u>* for another year.

TABLE V
INPUT SENTENCES THAT THE SYSTEM FAILED TO PARSE DUE TO A LACK OF KNOWLEDGE OF ARGUMENT STRUCTURES OF PREDICATES

(1) verb {+ object NP} {+ {*by* +} Extent NP} {+ *from* + Source NP} {+ *to* + Goal NP}	(a) *Increases each year since then have <u>lifted the annual rate to $2.08</u>.* (b) *Shares <u>rose 3.2 percent to $3.60</u> in early trading Tuesday.* (c) *Cash <u>increases by 18% from prior quarter to $2.6 billion</u>*
(2) deverbal noun + *of* + object NP + *to* + Goal NP	*a previous <u>projection of $2.9 billion to $3.04 billion</u>*
(3) verb governing *as*-phrase	*The carrier is also <u>using</u> its network infrastructure <u>as</u> collateral for financing*
(4) *leave/keep/hold* + NP + complement	(a) *Barclays says the stock's pullback from its 2015 high has <u>left it undervalued</u> relative to expected per-share profit growth of 13% this year and 13% in 2017.* (b) *That <u>kept</u> Exxon <u>in the black</u> last year, …* (c) *… <u>got you out of the market</u> before the biggest crashes*
(5) *remain* + complement	*But ExxonMobil (symbol XOM, $84.12) has <u>remained profitable</u>*

As the number of verbs that can appear in this pattern is limited, and the meanings of the verbs differ from those in different patterns as in *left the package at the door* (left a physical object at a place), the parser should be aware that verb meanings change according to the argument structures of predicates.

A note is in order. When the object NP complement of predicates in question was an adjective or a verb, the system was able to recognize the meaning of *keep* as in *keep the air fresh* or *keep us going*. However, when the complement was a PP, the reference to the formal structure only was not sufficient. For instance, both verb phrases *kept Exxon in the black* in (4b) and *kept it in a safe* have the same pattern: V + object NP + PP, and yet the meanings of *keep* in each pattern differ. It requires the system's reference to semantics to determine the meaning of *keep*, i.e. whether the object NP of the PP refers to an abstract entity like *in the black* or a concrete entity like *in a safe*.

F. Conjoined Constituents

As has long been discussed in the literature, to parse conjoined constituents is one of the most challenging tasks in NLP. Recent approaches have been statistical [11], [19], [22], [27]. Our investigation has detected many errors involving conjunction. Table VI illustrates such errors. Conjoined grammatical units are such as PPs, NPs, VPs, clauses, VP*ing*, and VP*en*.

A minimum requirement for a proper treatment of conjoined constituents is the same as for other constructions. Namely, the system should recognize the boundaries for constituents, whether they are clauses, VPs, NPs, or PPs. In addition to such syntactic knowledge, however, reference to simple semantic

[6] *To-PP* stands for a prepositional phrase (PP) beginning with the preposition *to*.

features of the conjoined constituents such as abstract and concrete entities will be helpful, since conjoined constituents are not only the same syntactic categories, but similar in simple semantic features [22].

TABLE VI
COORDINATE STRUCTURES THAT WERE NOT RECOGNIZED

(1) *to*-PP, *and nearly all the way back* + *to*-PP	*A substantial rebound in the market has taken the Dow Jones industrial average up to its highs of the year, and nearly all the way back to a level that hasn't been seen since June: 18,000.*
(2) NP, NP *and* NP modified by a relative clause	*…, delivering health services in a variety of ways, including via 9,600 retail pharmacies, 1,100 walk-in medical clinics and a drug-benefits manager (Caremark) that serves 75 million plan members*
(3) *both* NP *and* NP	*Honeywell's appetite for expansion presents both opportunity and risk for shareholders.*
(4) *from*-PP *and from*-PP	*Wal-Mart is battling intense price competition from dollar stores at the brick-and-mortar level and from Amazon.com (symbol AMZN) in cyberspace.*
(5) VP*ing and* VP*ing*	*Action Alerts PLUS, a charitable trust co-managed by Jim Cramer and Research Director Jack Mohr that is long Apple, has kept the same tact that they always have, characterizing this news as a small hiccup in Apple's long-term growth and remaining bullish on Apple's prospects.*
(6) *an ability to* VP *and* VP	*…, but has consistently demonstrated an ability to overcome near-term obstacles and create shareholder return.*
(7) *decide to* VP *and* VP	*… for Real Money Pro's Doug Kass, … to decide to lighten his Apple load, and short more of his shares premarket Thursday.*
(8) VP, VP *and* VP	*The promotions highlight a go-for-broke strategy by Chief Executive Officer Marcelo Claure, who is trying to cut $2.5 billion in costs, improve the network and add customers in a maturing wireless market.*
(9) modal verb + VP *and* VP	*… and those who respect it should stay on the sidelines or at least keep plenty of money in cash.*
(10) (subject) [VP*ing* and VP*ing*] , + (predicate) [V*ed*[7] … and V*ed* …]	*Cutting your stock-market exposure when the 50-day average fell below the 200-day average, and raising it when the 50-day rose above the 200-day, got you out of the market before the biggest crashes and kept you in it during the biggest bull markets.*

G.Ellipses

It is difficult to find an omitted element in an elliptical sentence, because the detection requires knowledge about what a full sentence is, and probably information about the preceding sentences, clauses or verb phrases as well. The finding task turns out even more difficult when a parenthetical phrase such as *it turns out* occurs in an elliptical clause, as in (1) of Table VII.

Ellipses are problematic in translation when the source and target languages differ in the rules of omission. With regards to the translation of elliptical sentences between English and Japanese, which are respectively a head-initial language and a head-final language,[8] there seems to be no single or general way to handle them, and further linguistic research in this regard remains to be conducted.

What will happen if a MT system fails to recognize an ellipsis? A case in point is following. The system in question treated an elliptical sentence as a full sentence. The clause *as it has so far this year* in (3) is an ellipsis. The full sentence is *as it has so far continued to do so this year*. The helping verb *have* was treated as the verb of possession, resulting in a gibberish output sentence.

TABLE VII
ELLIPSES THE SYSTEM FAILED TO RECOGNIZE

(1) *"Trees don't grow to the sky," the old Wall Street line goes. And neither, it turns out, do iPhone sales.*
(2) *No other country has displayed that. Not even the U.S.*
(3) *If the dollar continues to weaken against many foreign currencies, as it has so far this year, 3M's bottom line should benefit.*

H.Appositions

A large number of parsing failures occurred to noun phrases (NPs) containing appositions. NPs in (1-9) of Table VIII were not parsed correctly, triggering further parsing errors. Some appositions have complex structures, as in (1) and (6), while others simple structures. Even those having simple structures were not treated properly. NPs with long appositions were found frequently in the texts of newspapers and magazines. While problems of dealing with them remain to be solved, it is particularly important here for a parser to identify NPs.

TABLE VIII
NPS CONTAINING APPOSITIONS THE SYSTEM FAILED TO RECOGNIZE

(1) *And a key reason for the shortfall likely will be the first-ever drop in sales of iPhones, the iconic product that accounted for 66% of Apple's total sales of $234 billion in the fiscal year that ended last September.*
(2) *CVS's appeal is long-term income growth, not current yield.*
(3) *The good news for investors who want an above-average high yield is that Verizon's payout, currently an annual rate of $2.26 per share, appears safe.*
(4) *The most recent increase, an 8.3% boost this month, lifted the payout to an annualized $4.44 per share.*
(5) *But ExxonMobil (symbol XOM, $84.12) has remained profitable -- just not at its usual high levels.*
(6) *The 50 blue-chip companies in the group have raised their dividends every year for at least 25 years -- a record that makes the companies natural favorites of many retiree investors.*
(7) *One encouraging sign is that sales at Wal-Mart's U.S. stores open at least a year -- an important measure for assessing retailers -- have risen for six straight quarters.*
(8) *The carrier is also using its network infrastructure as collateral for financing -- an alternative to higher rates in the high-yield bond market.*
(9) *Honeywell's three divisions -- and particularly its aerospace unit --*

I. Floating Words and Phrases

An adverb or an adverbial phrase is allowed to occur at many locations in a sentence. Its freedom of location often caused problems. All the examples in Table IX contain an adverb or an adverbial phrase at unexpected positions. The adverb *long* before *been* in (1) caused a failure to recognize the phrase *has been the glaring exception.*

In (2), *over time* is between *the ability* and *to tweak technology…*, which are usually contiguous with each other, and the interference prevented the system from recognizing the *ability-to*-VP phrase. A similar example is (3), in which the phrase *in part* occurs between *thanks* and a *to*-PP, and obstructed the interpretation of the *thanks-to* phrase. Arrangements must be made for allowing an adverb or an adverbial phrase to appear even inside idiomatic phrases.

The system failed to parse examples in (4-6) because of an unexpected word or phrase that interferes with regular

[7] *Ved* stands for the past form of a verb.
[8] A head-initial language is a language in which governing elements such as the predicate verb of a sentence precede governed elements such as the object NP of the predicate verb. A head-final language is a language in which governing elements follow governed elements.

sequences of words: an adverb such as *repeatedly* appearing before the VP*ing* in (4), an adverbial phrase *to my surprise* occurring between *found* and its object *that*-clause in (5), and a parenthetical phrase *I think* occurring between *and* and a noun phrase *kind of a frightening one* even without parenthesizing *I think* with commas in (6). All these instances suggest that English grammar is fairly tolerant about the insertion of an adverb, an adverbial phrase, or a parenthetical phrase. This flexibility multiplies the number of possible strings of words. To deal with this proliferation, there should be rules to handle such floating elements, because they are fairly restricted to where to occur: they are likely to appear between phrases, but not within phrases. This general tendency should be captured. This also suggests the importance of the recognition of grammatical phrases.

TABLE IX
FLOATING ADVERBS, ADVERBIAL PHRASES, AND PARENTHETICAL PHRASES THAT CAUSED SYNTACTIC ERRORS

(1) 3M (symbol MMM, $164.91) has long been the glaring exception to the rule that a company should do one thing and do it well.
(2) 3M's genius lies in its ability over time to tweak technology to serve specific needs, …
(3) on fears that iPhone sales could be lighter in the coming quarters thanks in part to falling demand from China and emerging economies
(4) [clause], repeatedly coming up with new niche products across its four main sectors of abrasives, adhesives, coatings and filters.
(5) …, and found to my surprise that it's been a lot more useful than I had imagined.
(6) The Chinese stock market is a fairly remarkable phenomenon and I think kind of a frightening one.

V. CONCLUSION AND FUTURE WORK

We have seen a variety of sentence constructions the system failed to recognize. To deal with idiomatic constructions as well as MWEs requires knowledge about them as well as the ability to parse sentences into grammatical phrases and clauses. As the number of constructions to deal with is enormous, it would be practical to focus on and prepare knowledge about domain-specific constructions.

To solve syntactic problems such as coordinate structures, ellipses, and appositions requires a robust basic parsing ability, supported by the knowlege about argument structures of predicates as well as a simple semantic feature system as to whether the entity referred to is abstract or concrete. As observed, such elements as an adverb, an adverbial phrase, and a parenthetical phrase can intervene between phrases or between clauses, knowledge about clauses and phases (namely, what constitutes a clause or a phrase) is essential. Without such knowledge and grammatical rules, the number of strings of words we should prepare will increase exponentially, and eventually will be unmanageable.

As it is almost impossible to provide all the necessary knowledge to parse a great variety of sentence constructions, it will be practical to focus at first on knowledge and grammar to deal with domain-specific texts. In addition, some efforts should be made as to how to cope with unknown words and constructions, probably using statistics when a parser fails. There should be means that the system still should be able to come up with at least simplified versions of translations without losing the core meanings, even when a parser fails. Simplification, however, needs to recognize what is the head of a phrase or a clause. What is required urgently is a parser able to recognize a verb phrase and a noun phrase before identifying a clause, because this basic ability is not yet satisfactory at present.

REFERENCES

[1] Baldwin, T., Bannard, C., Tanaka, T. and Widdows, D. 2003. An Empirical Model of Multiword Expression Decomposability. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions Analysis, Acquisition and Treatment*. 89-96.
[2] Bunt, H. and A Van Horck. 1996. *Discontinuous Constituency*. Mouton De Gruyter.
[3] Elliot, D, Hartley, A., and Atwell, E. 2004. A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation. *AMTA 2004*. Pages 64-73.
[4] Church, K. 2013. How Many Multiword Expressions Do People Know? *ACM Transactions on Speech and Language Processing*. 10(2), Article 4: 1-13.
[5] Fillmore, C. J., Kay, P., and O'Connor, M. C. 1988. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language*, 64 (3), 501-538.
[6] Farrús, M., Costa-jussa, M. R., Marino, J. B., Posh, M., Hernandez, A., Henriquez, C., and Fonollosa, J. A. R. 2011. Overcoming statistical machine translation limitations: error analysis and proposed solutions for the Catalan-Spanish language pair. *Language Resources and Evaluation* (Springer). Vol. 45 Issue 2.
[7] Flanagan, M. 1994. Error classification for MT evaluation. *AMTA 1994*. 65-72.
[8] Goldberg, A. E. 1995. *A Construction Grammar Approach to Argument Structure*. Chicago: The University of Chicago Press.
[9] Goldberg, A. E. 2006. Constructions at Work: The Nature of Generalization in Language. Oxford: Oxford University Press.
[10] Hilpert, M. 2014. *Construction Grammar and its Application to English*. Edinburgh: Edinburgh University Press.
[11] Hogan, D. 2007. Coordinate Noun Phrase Disambiguation in a Generative Parsing Model. *Proc. of the 45th Annual Meeting of the ACL*, 680-687.
[12] Hunston, S. and Francis, G. 2000. *Pattern Grammar A corpus-driven approach to the lexical grammar of English*. Benjamins Publishing Co.
[13] Hurskainen, A. 2008 Multiword Expressions and Machine Translation. Technical Reports in Language Technology Report No 1, 2008. http://www.njas.helsinki.fi/salama/multiword-expressions-and-machine-translation.pdf (access date: Nov. 28, 2016).
[14] Kim, S. and Baldwin, T. 2013. Word Sense and Semantic Relations in Noun Compounds. *ACM Transactions on Speech and Language Processing*. 10(3), Article 9: 1-17.
[15] Kordoni, V. and Simova, I. 2014. Multiword Expressions in Machine Translation. LREC 2014. 1208-1211.
[16] Lau, J., Baldwin, T., and Hewman, D. 2013. On Collocations and Topic Models. *ACM Transactions on Speech and Language Processing*. 10(3), Article 10: 1-14.
[17] Metzler, D. P., Haas, S. W., and Cosic, C. L. Conjuction, Ellipsis, and Other Discontinuous Constituents in the Constituent Object Parser. *Information Processing & Management*, 26 (1): 53-71.
[18] Nadeau, D. and Sekine, S. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*. 30(1):3-26.
[19] Petrov, S and McDonald, R. 2012. Overview of the 2012 Shared Task on Parsing the Web. Notes of the First Workshop on Syntactic Analysis.
[20] Popović, M. and Burchardt, A. 2011. From Human to Automatic Error Classification for Machine Translation Output. *Proceedings of the 15th Conference of the European Association for Machine Translation*. 265-272.
[21] Ramisch, C., Villavicencio, A., and Kordoni, V. 2013. Introduction to the Special Issue on Multiword Expressions: From Theory to Practice and Use. *ACM Transactions on Speech and Language Processing*. 10(2), Article 3: 1-10.
[22] Resnik, P. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Jr. of Artificial Intelligence Research* 11. 95-130.

[23] Sag, I., Baldwin, T., Bond, F., Copestake, A, and Flickinger, D. 2002. Multiword Expressions: A Pain in the Neck for NLP, In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.
[24] Shutova, E., Kaplan, J., Teufel, S., and Korhonen, A. 2013. A Computational Model of Logical Metonymy. *ACM Transactions on Speech and Language Processing*. 10(3), Article 11:1-2.
[25] Stymne, S. and Ahrenberg, L. 2012. On the practice of error analysis for machine translation evaluation *LREC 2012*.
[26] Vilar, D., Xu, J., D'Haro, L., and Ney, H. 2006. Error Analysis of Statistical Machine Translation Output. *Proceedings of the LREC*. 697-702.
[27] Yoshimoto, A, Hara, K., Shimbo, M., Matsumoto, Y. 2015. Coordination-aware Dependency Parsing (Preliminary Report) *Proc. Of the 14th International Conference on Parsing Technologies*, pages 66-70.
[28] Google Language Tools at https://translate.google.com/ access dates: March-April, 2016.
[29] Alam, Y. 2017. Knowledge Required for Avoiding Lexical Errors at Machine Translation. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, February 2017, 7 pages.