[From: *Current issues in computational linguistics: in honour of Don Walker*, ed. Antonio Zampolli, Nicoletta Calzolari, Martha Palmer (Linguistica Computazionale, vol. 9-10); Pisa, Dordrecht, [1994]

Construction-Based MT Lexicons

Lori Levin
Sergei Nirenburg
Carnegie Mellon University
e-mail: {Isl, sergei}@nl.cs.cmu.edu

Abstract

This paper presents a novel view of the boundary between the generalizable and the idiosyncratic in MT lexicons. We argue that the domain of the idiosyncratic should, in fact, be broader than in most current approaches. While at present most MT systems involve phrasal lexicons, these typically contain terminology from a particular field. In order to facilitate naturalness of translation, specifically, to carry the level of “conventionality” of meaning expression across languages, it becomes necessary to use the concept of a grammatical construction, a (possibly, discontiguous) syntactic structure or productive syntactic pattern whose meaning it is often impossible to derive solely based on the meanings of its components. Identification of constructions allows an MT system to select the most appropriate conventional way of expressing a meaning from among the available ways. After discussing the notion of construction, we suggest the format for a construction lexicon for a knowledge-based MT system.

1 Introduction

Source text analyzers and target text generators in all rule-based MT systems rely on a variety of knowledge sources, centrally including grammars and lexicons. Grammar rules vary in generality (productivity) in that they can apply to very broad class of phenomena, such as, for instance, all adjectives, or to a very narrow class, even a single lexical unit. Information in lexicons typically relates to a single lexical entry (which, however, can be phrasal), but modern computational lexicons are often organized in such a way that some information applies to (often, broad) classes of entries.

There are many reasons to attempt to write grammar rules in the most general manner—the more generally applicable the rules, the fewer rules need to be written; the smaller the set of rules (of a given complexity) can be found to be sufficient for a particular task, the more elegant the solution, etc. In the area of the lexicon, the ideal of generalizability and productivity is to devise simple entries which, when used as data by a set of syntactic and semantic analysis operations, regularly yield predictable results in a compositional manner. To be maximally general, much of the information in lexical entries should be inherited based on class membership or should be predictable from general principles.

However, experience with NLP applications shows that the pursuit of generalization promises only limited success. In a multitude of routine cases it becomes difficult to use general rules. This leads to the necessity of directly recording (usually, in the lexicon) information about how to process small classes of phenomena which could not be covered by general rules. An important goal for developers of NLP systems is, thus, to find the correct balance between what can be processed on general principles and what is idiosyncratic in language, what we can calculate and what we have to know literally, what is compositional and what is conventional. In other words, the decision as to what to put into a set of general rules and what to store in a static knowledge base such as the lexicon becomes a crucial early decision in designing computational-linguistic theories and applications.

The trade-off between generality and idiosyncraticity is complicated by the possibility of the following compromise. If direct generalizations cannot be made, there may still be a possibility that the apparent variability in grammar rules and lexicon data can be accounted for by *parameterization:* there may exist a set of universal parameters that would explain the differences among various phenomena in terms of the difference in particular parameter settings. This is better than dealing with ungeneralized material. But the search for a set of universal parameters, however important, does not, in our opinion, hold a very bright promise from the standpoint of coverage.

In this paper we discuss a set of phenomena, *constructions,* which do not lend themselves to either a generalized or parameterized treatment, while at the same time differing from the predominantly, terminological material typically appearing in the phrasal parts of NLP lexicons.

This paper does not address the issue of how constructions can be actually used in the process of extracting meaning from texts, though in our long-term research program this is a central issue. In a nutshell, our approach to natural language analysis is built on a constellation of "microtheories" of particular language phenomena, united at the level of system architecture and representation language. The microtheory rules for calculating various components of meaning use a variety of diverse clues—morphological, syntactic, lexical—in their input conditions. We believe that knowledge about constructions—significant combinations of syntactic structures, lexical items and other morphemes—contributes greatly to the expressive power of such rules.

2 Idiosyncractic Phenomena in MT

The concept of construction was revived by Charles Fillmore, Paul Kay and their students (e.g., Fillmore et al., 1988, Fillmore and Kay, unpublished). According to this approach, the specification of a construction can include syntactic, semantic, and pragmatic information, but the semantics and/or pragmatics can be different from the compositional semantics and/or pragmatics normally associated with the same structure by productive rules. Furthermore, many constructions, such as those below, violate do not conform to general syntactic rules. Constructions are, therefore, like words in that they have to be learned separately as integral facts about language. At the same time, constructions are not the same as frozen idioms; they can be productive grammatical patterns, many of whose properties are predictable from general principles. The following are some examples of English and Russian constructions. The English examples are taken from Fillmore et al. (1988).

(1) a. What with the kids off to school and all.

b. Why not fix it yourself?

c. Him be a doctor?

d. What do you say we stop here?

e. It's time you brushed your teeth.

f. One more and I'll leave.

g. No writing on the walls!

(2) a.

Chto	ni	govori
what	CLITIC	say-IMPERATIVE

a	matematika	interesnyj	predmet
CONJ	mathematics	interesting	object

“Whatever you might say, mathematics is an interesting subject.”

b.

Chto	zhe	eto	ty
What	CLITIC	this-NEUTER	you-NOMINATIVE

Ivan Ivanovich,	zabyl	o
Ivan Ivanovich,	forgot	about

nashem	dogovore?
our	agreement-PREPOSITIONAL

“How come you forgot about our agreement, Ivan Ivanovich?”

c.

Kuda	nam	do	nix.
Where-DIRECTION	we-DATIVE	to	they-GENITIVE

“How can we compete/compare with them.”

Constructions with non-compositional semantics and pragmatics are not rare exceptions to rules. They co-exist with the basic lexis and grammar of language and in many cases present the most typical, unmarked way for expression of a particular meaning. Thus, the rules governing the use of constructions and the "regular" rules must be made to co-exist in any application, as they are equally important for associating semantic and pragmatic effects with utterances.

2.1 Conventionality and Constructional Divergences in MT

The *conventionality* of constructions is one of their defining characteristics. Constructions can be understood as conventional in two ways. First, meanings are associated with constructions by convention. That is, many constructions, similarly to separate words, have an arbitrary (non-compositional, non-iconic) association with their meanings. Among the types of meaning often associated with constructions are aspect, time/tense, modality, evidentiality, speaker attitude, speech act, conditionality, comparison, causality, rhetorical relations, etc. The microtheory approach to these phenomena makes abundant use of constructions as sources of information about what meanings are present in a sentence.

The other sense of conventionality concerns the typical default ways of expressing meanings in language, which may have been grammaticalized as an arbitrary form-meaning relationship. It should not be necessary to involve inference processes for the analyzer to arrive at the intended meaning. For example, (3)a is a conventional way of requesting that someone pass the salt, whereas (3)b is not a conventional request We propose different treatments for (3)a and (3)b. The former matches an entry in a *construction lexicon* and does not require inference. The latter requires inference in order to be interpreted as a request. ((3)a could also involve inference based on felicity conditions for requests, but its conventionality eliminates the need for inference.) All constructions in our lexicon are conventional in the first sense (that is, non-compositional but not all of them are conventional in the second sense. That is, in some cases a different realization of the same meaning could be less marked.

(3) a. Can you pass the salt?

b. Gee, this food is bland.

Conventionality is an important factor in translation. Thus, conventional expressions of a meaning in the source language should be translated into similarly conventional expressions of the same meaning in the target language. For example, the Japanese sentences in (4) are conventional expressions of the modal meaning of obligation and should therefore be translated into a conventional expression of obligation in English such as *You should go* or *You 'd better go.* Literal translations of these sentences into English. *(Not going won't do* and *The alternative that (you) went is good),* while understandable. are not appropriate translations, due to their low rating on the conventionality scale.

(4) a. Itta hoo ga ii.
go-PAST alternative NOM good
Literally: The alternative that (you) went is good.

b. Ikanakute wa ikenai.
go-NEG-GERUND TOP won’t do
Literally: Not going won’t do.

c. You’d better go./ You should go.

In translation, source text elements can vary in their degree of conventionality in expressing the meanings they carry. It is natural to require that the translation has the conventionality level closest to that of the source text. The requirement of maintaining conventionality levels in translation is a source of *constructional divergences,* instances (such as (4)) in which, in order to retain the level of conventionality, a target language passage is selected with a syntactic structure very different from that of the corresponding source passage. Further examples of constructional divergences are shown in (5)-(7).

(5) a. Ich esse gern.
I eat likingly
Literally: I eat likingly

b. I like eating. (Dorr, 1992)

(6) a. Juan suele ir a casa.
Juan be-in-the-habit-of go-INF to house
Literally: Juan tends to go home.

b. John usually goes home. (Dorr, 1992)

(7) a. On zagovoril.
he speak-lNCHOATlVE
‘He started to speak’

b. He started to read a book.

Examples like these, when discussed in the literature, have led to the impression that constructional divergences arise from a fairly limited set of correspondences (such as a syntactic head in one language corresponding to a modifier in another language). The Japanese examples in (4) and the Russian examples in (2) above show that this is not the case; the structures involved in constructional divergences can be radically different and, for the most part, do not appear to follow from predictable or regular correspondences between source and target languages.

2.2 What is the Impact of Constructions on MT?

Exploiting constructions seems to be the only way toward guaranteeing the production of conventional ("colloquial") rather than literal translations, although history of MT research shows neglect of the issue. Recent interest in MT divergences centers on linking and lexicalization divergences, which are less of a problem to solve, while largely ignoring the problem of constructions. It is clear that the expressive power of MT systems will grow significantly with the introduction of large construction lexicons.

3 Treatment of Constructional Divergences

A literal translation (that is, a translation which seeks to preserve in the target text the exact word and structure choice in the source text), even when formally possible, seldom succeeds in preserving the level of conventionality of the input text. Often a marked, unusual, compositional realization of the same meaning is obtained as illustrated in example (4). We believe that the treatment of constructional divergences in the lexicon is possible both for transfer-oriented and interlingual MT systems. Since our prior work in MT has centered mostly on the latter approach, our further discussion will be devoted to the ways of introducing the treatment of constructions into a lexicon structure similar to that used in some of our MT projects (e.g., DIANA, KBMT-89 or Mikrokosmos).

The text meaning representation produced for a construction will not, in the general case, be isomorphic to the syntactic representations of the source and target texts Nirenburg and Levin, 1992 for a discussion). This conclusion is further corroborated by the following consideration. Our theory of text meaning distinguishes between core semantic dependency statements (which we will call “the propositional content”) and additional semantic information that covers meanings such as aspect, time/tense, modality, evidentiality, speaker attitude, speech act, conditionality, comparison, rhetorical relations and others (which we will, for the sake of symmetry, call “non-propositional content” and which we represent as feature-value sets scoping over predicate-argument structures). The means to express these phenomena are among the most divergent among languages and at the same time are not readily parameterizable or generalizable.

Indeed, constructional divergences cannot be accounted for with a few parameters like head switching or locus of linking inside a semantic structure (Dorr, 1992). In our terms, constructions are used as a means of conventional, language-specific encoding of language-independent meaning. For example, the fact that the Japanese “`Sentence-past` *hoo ga ii*” conventionally encodes the meaning of suggestion or obligation is as much a part of the lexicon of Japanese as any definition of a word meaning. We introduce a *construction lexicon* as a repository of knowledge supporting both the treatment idiosyncratic, non-compositional constructions and the compositional realization of a variety of propositional meanings.

Before suggesting a possible structure of a construction lexicon entry, we would like to clarify a potential misunderstanding with respect to the definition of constructional divergences. It is important to distinguish constructional divergences from other circumstances that call for a target language translation to be structurally different from the source. For example, lexical gaps are typically treated in translation through optional, usually inferentially-produced paraphrases. For example, because there is a lexical gap for *afford* in Russian, a sentence like (8)a must be rendered in Russian as the translation of a sentence such as (8)b or (8)c. Examples in (4)-(7) are different in kind from the ones in (8) because in a computational implementation they should not involve paraphrasing through inference making but rather a look-up in a lexicon of conventional constructions (see below). Note that there are some indications that lexical gaps and constructional divergences form a scale rather than a dichotomy.

(8) a. John can’t afford a BMW.

b. John does not have enough money to buy a BMW.

c. John cannot allow himself to buy a BMW.

4 Some Examples

This section contains two examples of treatment of constructions in the framework of an interlingual MT system. In particular, it illustrates the lexicon entry structure and the interlingua (the text meaning representation, or TMR). The examples illustrate the use of constructions as units of analysis alongside words.

The examples also illustrate our treatment (or, rather, in our model, lack of the need for the treatment) of MT divergences—situations in which a source language sentence and its target language translation differ significantly in syntactic structure, syntactic category, or predicate-argument structure. No special mechanisms are needed to treat MT divergences in our model, as source and target language sentences are not expected to be isomorphic to the TMR or to each other. All that is needed in order to translate a sentence involving a divergence are source and target language lexical entries of the sort illustrated here that map different syntactic structures onto the same TMR.

For each example, we list a TMR, a source language syntactic structure and a target language syntactic structure. The languages used for illustration are English, Russian, and Japanese. (Since the system is symmetrical, we do not identify which is the source language and which is the target language in each example.) It should be obvious that the source and target language sentences can produce the same TMR even if their syntactic structures are not isomorphic.

The TMR structure consists of clauses which roughly correspond to the "who did what to whom" component of meaning but also includes information about speech acts, speaker attitudes, indices of the speech situation, and stylistic factors as well as relations (e.g., temporal ones) among any of the above, and other elements.

The examples also include the relevant zones of the source and target language lexical entries (namely, syntax and mapping to TMR). (This lexicon format is discussed in some detail by Meyer et al. 1990.) The first zone (`syn-struc`) specifies an LFG-style syntactic subcategorization frame (Bresnan, 1982) of a predicate including which grammatical functions (subject, object, complement, etc.) the predicate must appear with and any requirements the predicate has of those functions (case, syntactic category, specific lexical items, etc.).

The second lexical entry zone that we illustrate (`sem`) specifies the portion of TMR that is associated with the lexical item in question and how the components of the TMR correspond to the components of the syntactic zone. We have chosen examples in which the TMR is not isomorphic to the syntactic zone. In most of the examples, a complements of the lexical item heads the associated TMR. In these cases, the syntactic head of the sentence corresponds to a scope-taking operator or a simple feature-value pair in TMR.

4.1 Treating Speech Act Constructions

Consider the sentences in (9), which constitute conventional ways of making requests in Japanese and English. The TMR for both the English and the Japanese phrase is represented in Figure 1.

The TMR indicates that the speaker is performing the speech act of requesting (`speech-act-1`), that the request is that the hearer buy a book (`clause-1`), that the buying will occur after the time of speech (`relation-1`), and that some time after the buying (`relation-3`), the book will belong to the speaker (`relation-2`). The syntactic feature structures of the sentences are illustrated in Figure 2. The constituent structure of the Japanese sentence is presumably mono-clausal, but corresponds to the bi-clausal feature structure shown here. It is also possible that the feature structure should be tri-clausal, depending on the analysis of the potential morpheme. See Matsumoto (1992) for a recent discussion of these issues.

```
clause-1
  head:      buy-1
  agent:     *hearer*
  theme:     book-1

aspect:
  phase:     none
  duration:  momentary
  iteration: single

speech-act-1
  type:       request-action
  scope:      clause-1
  speaker:    *speaker*
  hearer:     *hearer*

relation- 1
  type:       temporal-before
  from:       time-of-speech
  to:         time-of(clause-1)

relation-2
  type:       possession
  from:       *speaker*
  to:         book-1

relation-3
  type:       temporal-before
  from:       time-of(clause-1)
  to:         time-of(relation-2)
```

Figure 1. TMR for the Sentences in Example (9).

(9) a. Hon o katte
book OBJ buy-GERUND

moraemasen ka?
receive-POTENTIAL-FORMAL-NEG QUEST

"Can I receive the favor of you buying a book for me?"

b. Can you buy me a book?

Figures 3, 4 and 5 show some of the the lexicon entries involved in building the syntactic structures shown in Figure 2 and the TMR shown in Figure 1.

The entries specify a) portions of the syntactic structure in which the lexical unit in question will appear, b) the meaning of the lexical unit, which could be viewed as "canned" portions of the TMR and c) correspondences between these two kinds of structure. The entries for *can* and *morau* (receive) are examples of construction lexicon entries. They contain information that is specific to the use of these words in making requests such as the subject being *you* in English and the tense being non-past in

```
English Syntactic Structure

    PREDICATE           can
    SUBJECT             you
    COMPLEMENT
           PREDICATE    buy
           SUBJECT      you
           OBJECT       me
           OBJECT2      book

Japanese Syntactic Structure

   PREDICATE           morau (receive)
   MOOD                potential, interrogative
   TENSE               non-past
   OBJECT              hon (book)
   SUBJECT             pronoun (speaker)
   COMPLEMENT
           PREDICATE   kau (buy)
           SUBJECT     pronoun (hearer)
           OBJECT      hon (book)
```

Figure 2. Syntactic Structures for the Sentences in Example (10).

```
CAN:

Syn-Struc:  predicate: [0] can
            subject:  [1]
               root: you
            complement: [2]
               subject: [1]

Sem:        clause [3]
               head: meaning-of ([2])
            speech-act [4]
            type: request-action
            scope: [3]
            speaker: *speaker*
            hearer: *hearer*
            relation [5]
               type: temporal-before
               from: time-of-speech
               to:   time-of([3])
```

Figure 3. Construction Lexicon Entry for a Request in English.

Japanese. *Can* and *morau* have other lexical entries as well for their other senses and constructional uses. There are also many other construction lexicon entries for other ways of making requests in both languages.

```
BUY

Syn-Struc:  predicate: [0] buy
             subject: [1]
             object: [2]
             object-2: [3]

Sem:        clause [4]
               head: meaning-of([0])
               agent: meaning-of ([!])
               theme: meaning-of ([3])
             relation [5]
               type: possession
               from: meaning-of([2])
               to: meaning-of([3 ])
             relation [6]
               type: temporal-before
               from: time-off[4])
               to:  time-of([5])
```

Figure 4. Lexicon entry for English “buy.”

```
MORAU

Syn-Struc:  predicate:  [0] morau
             tense: non-past
             mood: potential
             subject:  [1]
                root: speaker
             object: [2]
             object-2:  [3]
                root: hearer
             complement: [4]
                inflection: gerund
                subject: [3]
                object:  [2]

Sem:        clause [5]
               head: meaning-of ([4])
             speech-act [6]
                type: request-action
                scope:  [5]
                speaker: *speaker*
               hearer: *hearer*
             relation [7]
                type: temporal-before
                from: time-of-speech
                to:  time-of ([5])
```

Figure 5. Construction Lexicon Entry for a Request in Japanese.

Correspondences between syntactic elements in the syn-struc zone and elements of TMR in the sem zone are indicated by co-indexing. Meaning-of (x) is a function whose value is the TMR corresponding to the feature structure with index x. Notice that in the lexical entries for *can* and *morau* the complement of the syn-struc zone is co-indexed with the head of the main clause in the sem zone. In building a TMR for the English sentence in (9), this means that the meaning-of the syntactic clause headed by *buy,* given in the sem zone of Figure 4, will become the head of the main clause of the TMR in Figure 1. In other words, *buy* (or *kau)* conveys the main semantic content of the sentence. *Can* and *morau* serve only to trigger the speech-act request-action in the TMR.

4.2 Treating Modality

Our second example involves constructions that illustrate the modality of obligation in Japanese, Russian, and English, as shown in (10). The TMR corresponding to these sentences is shown in Figure 6.

```
clause-l
  head:            go-1
  agent:           *hearer*
  destination:     *unknown*
  aspect:
      phase:       none
      duration:    *unspecified*
      iteration:   single

attitude-1
  type:             deontic
  value:            0.8-1.0
  scope:            clause-l
  attributed-to:    *speaker*
  time:             time-of-speech

relation-1
  type:              temporal-before
  from:              time-of-speech
  to:                time-of(clause-l)
```

Figure 6. TMR for the sentences in (10).

The frame attitude-1 in the TMR indicates that the speaker has a fairly strong (value .8-1.0) deontic attitude toward clause-l, which says that the hearer goes somewhere unspecified. The syntactic structures for the English, Japanese, and Russian sentences in (10) are shown in Figure 7. These examples illustrate a constructional divergence in that the syntactic structures used by the individual languages to express the same meaning are radically different.

```
English Syntactic Structure

    PREDICATE          had better
    SUBJECT            you
    COMPLEMENT
           PREDICATE   go
           SUBJECT     you

Japanese Syntactic Structure

    PREDICATE           ii (good)
    SUBJECT             hoo (alternative)
          REL-CLAUSE
             PREDICATE  itta
             SUBJECT    pronoun

Russian Syntactic Structure

     PREDICATE           stoit'  (cost)
     SUBJECT             tebe (you-dative)
     COMPLEMENT
           PREDICATE     pojti (go)
           SUBJECT       tebe
```

Figure 7. Syntactic Structures for the Sentences in (10).

(10) a. You'd better go.

b. Itta hoo ga ii.
go-PAST alternative SUBJ good
"The alternative that you went is good."

c. Tebe stoit pojti.
you-DATIVE cost-IMPERSONAL go-INFINITIVE
"To you costs to go."

The lexicon entries in Figures 8,9 and 10 show how the different syntactic structures are mapped onto the same TMR. We have indexed them by their most salient lexical item. The syn-struc fields characterize the language-specific realization of the construction.

For example, the syn-struc field of the English example says that this construction is headed by a verb *had* occurring with the adverb *better*, which takes a noun phrase as a subject and an infinitival clause as a complement.

The Japanese syn-struc field for *hoo* says that this construction is headed by an adjective such as *ii* or *tanosii* which is predicated of the noun *hoo* which is, in turn, modified by a relative clause in the past tense.

```
HAD-BETTER

Syn-Struc:  predicate: [0]
            tense: past
            adverb: better
            subject: [1]
            complement: [2]
               subject: [1]
               inflection: infinitive

Sem:       clause  [3]
               head: meaning-of ([2])
            attitude [4]
               type: deontic
               value: 0.8-1.0
               scope: [3]
               attributed-to: *speaker*
               time: time-of-speech
            relation [5]
               type: temporal-before
               from: time-of-speech
               to:  time-of([3])
```

Figure 8. Construction Lexicon Entry for Expressing Deontic Modality in English.

In the Russian example (Figure 10), we are taking the verb *stoit'* to show impersonal agreement typical with the non-nominative subject *tebe*. Other analyses are possible. Again, it is important to note that the words *had, hoo,* and *stoit'* have other lexical entries corresponding to their other senses and uses. There are also many other construction lexicon entries corresponding to different constructions that also express deontic modality.

The `sem` fields in Figures 8,9 and 10 contain TMR templates, which will give rise the TMR shown in Figure 6 when filled in. The English `sem` field indicates that the meaning of the complement of *had better* will become the main clause of the TMR. Similarly, the meaning of the complement of *stoit'* will become the main clause of the TMR. The Japanese entry for *hoo* indicates that the relative clause attached to *hoo* will supply the main propositional content in the TMR. In all three `sem` fields there is an additional component of meaning that says that the proposition expresses a high positive level of the speaker's deontic attitude toward the content of the proposition. The coindexings between the `syn-struc` and `sem` zones of these lexical entries will result in the same TMR being built even though the syntactic structures of the constructions are markedly different.

The above examples certainly cannot take the place of a full theoretical specification of the use of construction and should be viewed as a set of pre-theoretical intuitive considerations on the basis of which exploratory system development will occur. Armed

```
HOO

Syn-Struc:  predicate: [1]
             root: (OR ii, tanosii. etc.)
             subject: [0]
                relative-clause: [2]
                   tense: past

Sem:        clause [4]
               head: meaning-of ([2])
            attitude [5]
              type: deontic
              value: 0.8-1.0
              scope: [4]
              attributed-to: *speaker*
              time: time-of-speech
            relation: [6]
              type: temporal-before
              from: time-of-speech
              time-of ([4])
```

Figure 9. Construction Lexicon Entry for Expressing Deontic Modality in Japanese.

with the results produced by such an exploratory prototype we will, in our future work. proceed to formulating a more strict statement about treatment of constructions in a multi-lingual environment.

5 The Passive: Principled or Conventional?

In suggesting that vast numbers of constructions should be represented as entries in a construction lexicon, we are not recommending that the principled and rule-based aspects of language be ignored. In fact, our model of MT explicitly allows us to represent both compositional and conventional aspects of each construction. We will use the English passive construction to illustrate the interaction of the compositional and the conventional.

In reaction to older rule-based theories of syntax, proponents of modern principle-based theories have implied that the many constructions are figments of our imaginations. That is, the constellation of syntactic structures that make up a construction are not a unified phenomenon, but an accidental co-occurrence of independent phenomena that are each predicted by general principles. Some of the principle-based phenomena involved in the English passive are listed below. (See, for example, Levin (1988), Bresnan and Kanerva (1989) and Marantz, 1984.) We have attempted to present them in neutral way that is applicable to a number of different syntactic theories. They include, among other things:

```
STOIT'

Syn-Struc:  predicate: [0] stoit'
            subject: [1]
            case: dative
            complement: [2]
            subject: [1]
            inflection: infinitive

Sem:        clause [3]
               head: meaning-of ([2])
            attitude [4]
               type: deontic
               value: 0.8-1.0
               scope: [3]
            attributed-to: *speaker*
            time: time-of-speech
            relation [5]
              type: temporal-before
              from: time-of-speech
              to:  time-of ([3])
```

Figure 10. Construction Lexicon Entry for Expressing Deontic Modality in Russian.

- Morphemes, such as the passive participle morpheme, that unlink an agent or external argument from its syntactic position are common cross-linguistically. This is predicted by theories of the interaction of morphology and syntax or by theories of grammatical relations. It leaves the subject position open so that it can be filled by something else.

- Some principle of grammar determines that a direct object can become a subject when the agent or external argument has been unlinked (and there is not a locative or expletive element in subject position). This allows the active verb's direct object to correspond to the subject of the passive verb.

- Because the English past participle is not tensed, it must occur with a tensed auxiliary verb when it is in a main clause or any other environment that requires a tensed verb.

The following examples show that these are in fact three separate components to the passive, each of which can occur independently given the right circumstances. Example (11) illustrates passives without *be* in environments that do not require a tensed verb or that include another tensed verb. Example (12) illustrates unlinking of the agent argument without promotion of object to subject when a locative or expletive element is in subject position. Example (13) illustrates promotion of a direct object without passive in other constructions that involve unlinking of an agent or external argument.

(11) a. *Admired by everyone,* she was sure to win the election.

b. They got *arrested by the police.*

c. We had them *arrested by the police.*

(12) a. In this spot, well toward the center and front of the vast herd, appeared about to be enacted *a battle between a monarch and his latest rival for supremacy.* (Zane Grey)

b. The wall paper was discoloured with age; it was dark grey, and there could be vaguely seen on it *garlands of brown leaves.* (W. Somerset Maugham)

c. Here, in the stone wall, had been wonderfully carved by wind or washed by water *several deep caves* above the level of the terrace. (Zane Grey)

d. Nowhere could be gotten *a better idea of its age* than in this gigantic silent tomb. (Zane Grey)

(13) a. The bread cut easily,

b. The glass broke.

In spite of these general, independent principles, there are strong reasons to view the passive as a unified construction for the purpose of machine translation. Although the components of the passive are each independently motivated, when they co-occur, they take on a range of meanings and functions that are not present when the components of the passive occur independently. The presence of the construction as a whole might, for example, signal certain interpretations of discourse focus or tense and aspect. These interpretations are neither inherent in nor unique to the passive construction, and may in fact require translation into different constructions in different target languages. Therefore, it is important to recognize the co-occurrence of the independent components so that specific meanings can be associated with the construction as a whole. In interlingual MT, those meanings should then be represented in the interlingua text in a way that is independent of the syntax of the English passive.

Processing of the passive construction can involve both a construction-based approach and a compositional syntactic analysis based on principles of syntactic theory. After a sentence has been parsed using compositional, theoretically motivated syntactic rules, the special co-occurrence of the independent components of the passive will be recognized by a construction lexicon entry such as the one in Figure 11.

The `sem` field of this entry indicates that the subject of the passive sentence is more salient than the oblique agent argument. It can also contain information related to other microtheories such as those of tense and aspect. This entry is indexed by the lexical item *be*. Other uses of passive verbs without *be* will be covered in separate lexical entries.

```
BE

Syn-Struc:  predicate: [0] be
            subject: [1]
            complement: [2]
            subject: [1]
            oblique: [3]
                  preposition: by
               inflection: past-participle
               voice: passive

Sem:        clause [4]
               head: meaning-of ([2])
            attitude [4]
               type: saliency
               value: . 4
               scope: meaning-of ([3]}
               attributed-to: *speaker*
               time: time-of-speech
            attitude [4]
               type: saliency
               value: . 6
               scope:   meaning-of   ([1])
               attributed-to:,*speaker*
               time:   time-of-speech
```

Figure 11. Construction Lexicon Entry for Passive Verbs with *Be.*

6 Conclusion

This paper presented a novel view of the boundary between the generalizable and the idiosyncratic in MT lexicons. We argue that the domain of the idiosyncratic should, in fact, be broader than in most current approaches. While at present most MT systems involve phrasal lexicons, these typically contain terminology from a particular field. In order to facilitate naturalness of translation, specifically, to carry the level of "conventionality" of meaning expression across languages, it becomes necessary to use the concept of a construction, a (possibly, discontiguous and productive) phrase whose meaning it is often impossible to derive solely based on the meanings of its components. It is also necessary to identify a construction in order to be able to select the most appropriate conventional way of expressing a meaning from among the available ways.

We discussed constructions in terms of the phenomenon of MT divergences. We have then shown how to incorporate the treatment of constructions into a standard interlingual MT environment, without losing syntactic or semantic generality of this approach. We claim also that treatment of constructions is both essential and attainable for the other major rule-based MT paradigm, the transfer approach.

References

[1] Bresnan, J., *The Mental Representation of Grammatical Relations.* MIT Press Cambridge, MA, 1982.

[2] Bresnan, J. and J. Kanerva, “Locative Inversion in Chichewa: A Case Study of Factorization in Grammar”, *Linguistic Inquiry,* Vol. 20, 1989, 1-50.

[3] Dorr, B., “Classification of Machine Translation Divergences and a Proposed Solution”, *Computational Linguistics,* 1992.

[4] Fillmore, C., P. Kay and M.C. O'Connor, “Regularity and Idiomaticity in Grammatical Constructions: The Case of *Let Alone”*, *Language,* 64, 1988, 501-38.

[5] Fillmore, C. and P. Kay, *Linguistics X20: Construction Grammar Coursework.* Chapters 1-11. Unpublished lecture notes. University of California at Berkeley.

[6] Levin, L.S., *Operations on Lexical Forms,* Garland, New York and London, 1988.

[7] Marantz, A., *On the Nature of Grammatical Relations,* MIT Press, Cambridge. MA, 1984.

[8] Matsumoto, Y. 1992. *On the Wordhood of Japanese Complex Predicates.* Ph.D. dissertation. Stanford University.

[9] Meyer, I., B. Onyshkevych and L. Carlson, “Lexicographic Principles and Design for Knowledge-Based Machine Translation”, CMU-CMT Technical Report 90-118, Center for Machine Translation, Carnegie Mellon University, 1990.

[10] Nirenburg, S. and L. Levin, “Syntax-Driven and Ontology-Driven Lexical Semantics”, in J. Pustejovsky and S. Bergler, (eds.), *Lexical Semantics and Knowledge Representation,* Springer-Verlag, 1992, 5-20.