Principles and Idiosyncracies in MT Lexicons

Lori Levin and Sergei Nirenburg

Center for Machine Translation
School of Computer Science
Carnegie Mellon University
{lsl, sergei}@nl.cs.cmu.edu

1. The Problem

The purpose of an MT lexicon is to facilitate translation. If we adhere to a symbolic (non-statistical) approach then we need a representation of meaning as the basis of translation. This representation is produced by programs typically called semantic and pragmatic interpreters, based on a variety of knowledge sources, such as grammars and lexicons. The grammar rules and lexicon entries vary in generality (productivity) in that they can apply to a very broad class of phenomena, such as all nouns, or to a particular lexical unit, such as *however*. The size of these rule domains forms, in fact, a range from the former to the latter. It is, naturally, preferable, from the point of view of generality of theory and economy ("capturing generalizations") to prefer the use of few powerful rules applicable to broad domains. However, language analysis shows that this noble pursuit promises but limited success. In a multitude of routine cases it becomes difficult to make generalizations. This leads to the necessity of directly recording information about how to process small classes of phenomena, specifically those that could not be covered by general rules. The trick is to find the watershed 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.

When an optimum balance on this issue is achieved, a lot of benefits immediately accrue. Thus, the maximum possible generality facilitates extensibility of description at minimum cost, while the presence of a large number of idiosyncratic descriptions guarantees breadth of coverage. Thus, the decision as to *what to put into the 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.

The situation becomes even more pronounced in the case of multilinguality. In order to maximize reusability of knowledge for the description of additional languages, it is preferable to factor out all the general rules that either directly hold across particular languages. Thus, an additional task is to try to determine *which phenomena carry across languages (and so can be treated in a general way) and which are language-specific.* (The answer will, of course, depend on a particular set of languages.)

Finally, 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 the next best thing to direct

generalization. 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 will investigate the balance between the productive, predictable aspects of language and the non-productive, idiosyncratic aspects of language viewed through the prism of the machine translation application. Specifically, we will look at phenomena that give rise to MT divergences. We will concentrate on three types of divergences involving lexicalization of complex meanings, linking semantic roles to syntactic argument positions and the use of syntactic constructions to express a given meaning.

2. MT Divergences

It is well documented in MT literature that languages differ in the realization of particular types of meaning. Among the differences in realization are:

- **Linking Divergences:** different mappings of semantic predicate-argument structure onto syntactic structure (how case roles map into syntactic arguments, etc. — *gustar* vs. *like* or the unavailability of the dative shift in, say, French or Spanish).
- **Lexicalization Divergences:** different mappings of the set of meaning components onto a set of lexical units of a language (this category subsumes the well-known example, due to Len Talmy, of manner conflation in the English *the bottle floated into the cave* which must be rendered in Spanish as *la botella entró a la cueva flotando*; these are *lexicalization* divergences;
- **Constructional divergences**, which are made manifest by the fact that different languages use completely different syntactic structures to realize a particular type of meaning (some examples were discussed in Dorr, 1992 and Arnold and des Tombe, 1987).

In the rest of this section we discuss the above classes of divergences. We concentrate on the extent to which they are parameterizable across languages and the plausibility of their being governed by general rules a particular language.

2.1. Linking and Lexicalization Divergences

A good perspective on lexicalization divergences can be obtained by considering the task of lexical selection in text generation. At input there is a text meaning representation which can be viewed, ultimately, as a large set of property-value pairs or meaning constraints. The process of lexical selection consists of finding target language lexical units that cover the set of these property-value pairs as densely as possible (cf., e.g., Nirenburg and Nirenburg, 1988).[1] Different languages will

[1]This is, of course, a simplification, as some of the input meanings will be realized not lexically at all but rather through syntactic constructions, word order, etc.

suggest different covers for the same material. We will illustrate three cases of lexicalization differences among languages.

Differences in meaning coverage of particular lexemes ((1)) have been abundantly discussed in linguistic (Talmy, 1985) and MT (e.g., Dorr, 1992) literature under the heading of "conflations."

(1)

a. La botella entró a la cueva flotando.
the bottle entered to the cave floating
Literally: The bottle entered the cave floating.

b. The bottle floated into the cave.

In the above example, Spanish does not allow for covering the meaning of "enter while floating" in a single word, whereas English does. This state of affairs seems to hold for all manner-of-motion verbs in English (Talmy, 1985) and therefore are candidates for processing by general rules which would be applicable to the entire class of manner-of-motion verbs. Similarly, German happens to have a single word, "Schimmel" which covers the meaning of "white horse" which in English will have to be expressed by a phrase. There is nothing that can be generalized in this instance, and therefore, this information should be considered known rather than calculated and thus stored in the lexicon.

There are other types of lexicalization divergences. One such involves the necessity to use different words in one language where in another the same word would do. For example, Russian has two, and Hebrew, three words corresponding to the English word "wear." The choice in Russian and Hebrew depends on what you are wearing:

English	Russian	Hebrew	
wear	nadet'	havash	(hat)
wear	nadet'	lavash	(coat)
wear	obut'	na'al	(shoes)

In any given line in the table below each language uses one word to express the meaning of "wear;" however, depending on the context, different words will be selected, due to cross-linguistic differences in selectional restrictions. It is difficult to think how this phenomenon can be governable by a general rule.

Another class of examples makes manifest differences in selection and use of semi-functional verbs across languages. Thus, in one language a conventional way of expressing a meaning could involve a construction with a semi-functional verb and a noun, while in another the same meaning can be realized through a verb only ((2)

(2)

a. I have difficulty ...

b. Ja zatrudnyajus' ...

The above realizations are primarily collocational and, as such, cannot be subject to general rules. This information needs to be stored in the lexicon.

Linking divergences stem from difference in the assignment of thematic roles to grammatical functions in different languages. For example, a Recipient can be linked to the grammatical function Object in English, whereas in Russian, for instance, Recipient must be an oblique object marked with dative case, as in Example (3).

(3)

a. Ja dal studentam knigi.
I gave-MASC students-DATIVE books-ACCUSATIVE.
Literally: I gave to-the-students books.

b. I gave the students books.

In Example (4) the experiencer is the subject in English but does not appear to be the subject in Spanish.

(4)

a. Me gustan las manzanas.
me like-3PL the apples.
Literally: me like the apples.

b. I like apples.

Many linking divergences are systematic and if there is hope for parameterization and generalized rules in solving divergence problems, this is the class of divergences where this could be most successful.

Our approach to linking and lexicalization divergences is predicated on what we perceive as a basic dichotomy in linguistic semantics – the difference between studying semantically-relevant information that is encoded in the syntax of particular languages (syntax-driven semantics) and extracting – and representing in a language-neutral way – semantic information encoded primarily in the lexis of a particular language, through the connections between the lexical items and elements of a world model, understood as universal semantic units (ontology-driven semantics). In a nutshell, syntax-driven lexical semantics identifies semantic classes of verbs that share subcategorization and thematic role assignment properties. Ontology-driven lexical semantics devises an inventory of properties and their value sets which describe the meanings of words and expressions in natural language. Our understanding of the purview of the two types of lexical semantics and their interrelationship is described in Nirenburg and Levin (1991). In brief, the task of deriving the propositional content of a unit of input will, in our theory, involve three levels of representation: grammatical structure, language-specific lexical semantics (based on universal principles with parameters of variation), and ontology-based language-independent text meaning.

The motivation for our theory of computational semantics is as follows. We see the syntax of particular languages as a unique coding system for semantic and pragmatic information. For example, the semantic case role AGENT can be syntactically coded using nouns in nominative

case that agree with the active-voice main verb in any or all of gender, number and person.[2] A major component of this coding system relates to representing predicates and their arguments. As each language has an idiosyncratic system of coding predicates and their arguments, we allow for an intermediate level of representation which we call the syntax-driven lexical-semantic level.

This representation relies on knowledge about equivalence classes of predicates defined by the way they code their arguments in each language. These lexical-semantic classes supply knowledge necessary for determining the propositional stratum of meaning by suggesting where in the syntactic structure to look for the arguments and the predicate. The lexical-semantic systems of various languages have a lot in common. But at the same time, they also systematically differ in a number of respects. For example, the classes of verbs undergoing particular transitivity alternations will be different in different languages. Proceeding from the simplifying assumption that, barring lexical gaps, every meaning can be expressed in every language, we seek to determine a level of meaning representation which does not feature variations among languages. Thus, our final meaning representation is motivated not by language-dependent facts but rather by an independent ontology or world model.

The differentiation of the two classes of lexical semantics leads to a solution of the linking and lexicalization divergence problem in the following way. Some of the classes of the above divergences – notably, linking and conflation – have a potential for being generalizable, that is we can avoid listing solutions to them in individual lexicon entries but rather rely on more or less general rules. The most general rules will be nearly universal; some others will abide by general parameters (for example, whether a language allows conflation of manner with change-of-location verbs); some others still will be applicable to a single language. Basically, though, the phenomena subject to divergences are language-dependent and are captured in their respective SDLSs. After the SDLS of each language has been used to decode the semantic role assignment in a particular input, an additional set of language-specific rules will map the lexical meanings of the source language words and phrases into a language-neutral form motivated by a particular ODLS. During generation, the process is essentially reversed.

It should be noted, however, that although linking is highly rule-governed and parametrizable, there are many idiosyncratic exceptions to these rules; lexicalization divergences, except for a few types of conflation, are not generalizable and must be treated individually in the lexicon.

2.2. Constructional Divergences

Work on divergences in interlingual machine translation has largely concentrated on identifying parameters of variation in linking interlingual concepts to syntactic configurations. While this has been largely fruitful for divergences in linking and lexicalization, we believe that this approach should not be applied to constructional divergences. Although the data on constructional divergences presented in most papers seems to indicate clean and systematic variation among languages, further examination of relevant data reveals drastic, unsystematic differences in how languages ex-

[2]There is, of course, more to the study of syntax than seeing it as a coding for semantic and pragmatic information. But this minimalist view of syntax will suffice for our purposes — analysis and generation of texts in a number of languages.

press the same meaning. We concluded that it would be inappropriate to continue to make attempts to explain the broad inventory of constructional divergences using principles and parameters.

2.2.1. Constructions

To defend our position effectively, it is necessary first to discuss the concept of construction (as revived by Fillmore et al., 1988) and the distinction between conventional ways of realizing meanings and (sometimes non-colloquial) free paraphrases which can be generated to create the same effects.

The approach to constructions taken by Fillmore et al. (1988) is based on the following premises. 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 syntactic structure by productive rules. Constructions are, therefore, like words in that they have to be learned separately as integral facts about pieces of the language. On the other hand, constructions are not necessarily frozen idioms; they can be productive grammatical patterns, many of whose properties are predictable from general principles.

Note that 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 offer the desirable option for expression of meaning. In fact, almost every sentence in every text involves at least one construction. 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.

The study of constructional divergences for the application of translation rests on the concept of conventionality in language. We claim that many constructions, like words, can 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 distinction between conventional and non-conventional expressions of meaning is not always clear-cut. However, when we talk about expressing some meaning conventionally, we refer to the usual, typical, "default" way of expressing this meaning in the 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, *You should go* in (5) is an instance of a conventional expression of deontic modality in English, whereas *Not going won't do* is less conventional and *The alternative that (you) went is good.* is thoroughly unconventional, though parsable. The less conventional a construction is the more difficult it is to process it.

(5)

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 should go.

2.2.2. Treatment of Constructional Divergences

In translation it is always desirable to render the conventional expression of a source language meaning into a conventional expression of the same meaning in the target language. So, for example, the Japanese example ((5)b) should be translated as ((5)c), not as *Not going won't do.*

We observe that 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 this respect. At best, one can expect to produce a marked, unusual, compositional realization of the same meaning. For instance, in (5) the literal English translation *Not going won't do* may convey the intended meaning, but certainly not in a conventional way, whereas the Japanese phrase is a conventional method of realizing the meaning. The much more preferable translation options, *You should go* or *You should go* are structurally very different from the Japanese sentence. Thus, if the preservation of the conventionality level of the source text is a basic goal in translation, one should be prepared to forgo the reliance on possible structural correspondences among source and target texts. The meaning representation which is the result of the analysis process will, thus, not be isomorphic to the SDLS-oriented representations of the source and target texts. This conclusion is further corroborated by the following consideration.

The SDLS/ODLS theory 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 a system based on the SDLS/ODLS will 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 not readily parameterizable or generalizable.

In fact, we posit that nonpropositional content gets represented in the language-independent text meaning formalism directly, bypassing the regular SDLS-to-ODLS linking rules. We introduce a *construction lexicon* as a repository of knowledge supporting both the mapping of idiosyncratic noncompositional constructions into feature-value sets of the language-independent meaning representation and the choice of conventional realization of a variety of propositional meanings in particular languages.

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.

Thus, there is a lexical gap for "afford" in Russian. Therefore, a sentence like (6)a must be rendered in Russian as the translation of a sentence such as (6)b or (6)c. Examples in (4) are different in kind from the ones in (5) 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).[3]

(6)

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.

Constructional divergences cannot be accounted for with a few parameters like head switching or locus of linking inside a semantic structure. 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 "S-past hoo ga ii" conventionally encodes the meaning of obligation is as much a part of the lexicon of Japanese as any definition of a word meaning.

3. Construction Lexicon: An Example

In practical machine translation systems it has been common practice to add phrase lexicon entries to the "regular" ones for efficiency reasons. Our approach combines the expected improvement in the efficiency of the system which uses the construction approach with a theoretically important contribution of the treatment of constructions. The representation format of the construction lexicon and actual mechanics of its use follow the method used in the DIANA NLP project (e.g., Meyer et al., 1990). Entries from the construction lexicons of English and Japanese follow.

```
SHOULD:   (SYN-STRUC
           ((root var0)
            (subj ((root var1) (cat n)))
            (xcomp ((root var2) (cat v)
                    (subj ((root var1)))
                    (vform infinitive)))))
          (SEM (LEX-MAP
             (proposition var3
                  (head (meaning-of (var1))))
             (SPEAKER-ATTITUDE
                  (ATTRIBUTED-TO *speaker*)
                  (SCOPE var3)
                  (TYPE deontic)
                  (VALUE (range 0.7 1))))))
```

[3]Note that there are some indications that the lexical gaps and the constructional divergences form a scale rather than a dichotomy.

```
HOO:      (SYN-STRUC
            ((subj ((root var0)
                    (rel-clause ((root var1)
                                 (tense past)))))
             (root var2)(value (*OR* ii tanosii))
             (cat adj)
             (tense non-past)))
          (SEM (LEX-MAP
             (proposition var3
                  (head (meaning-of (var1))))
             (SPEAKER-ATTITUDE
                  (ATTRIBUTED-TO *speaker*)
                  (SCOPE var3)
                  (TYPE deontic)
                  (VALUE (range 0.7 1)))))
```

The above examples show entries in the construction lexicons of English and Japanese for the constructions illustrated in (5)a and c. We have indexed them by their most salient lexical item. The SYN-STRUC fields are skeletal LFG-like f-structures characterizing the construction. For example, the SYN-STRUC field of the English example says that this construction is headed by a verb "should" which takes a subject which should be a noun phrase and a complement infinitive verb phrase. The Japanese SYN-STRUC field 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.

The SEM fields contain a template, parts of which are co-indexed to elements of the SYN-STRUC field. The English SEM field says that the head of the proposition expressing the meaning of this construction will be filled by the semantic interpretation of the complement of "should," whereas in the Japanese SEM field, the proposition head will be filled by the interpretation of the relative clause. In both 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. When these entries are used by a syntactic and semantic analyzer for processing the sentence "You should go" or "Itta hoo ga ii," the following language-neutral meaning representation gets produced. Even though the syntactic structures of the constructions are markedly different, they convey the same meaning.

```
(PROPOSITION-25
  (HEAD  (\%go-1))
  (AGENT *hearer*))

(SPEAKER-ATTITUDE-6
 (ATTRIBUTED-TO *speaker*)
 (SCOPE PROPOSITION-25)
 (TYPE deontic)
 (VALUE (range 0.7 1)))
```

4. Conclusion

To summarize, our position on the treatment of MT divergences is as follows. We realize that different types of divergences lend themselves to a varying degree to generalizations, and in those cases where this is possible, we should make use of general rules and parameters. Our lexical-semantic theory facilitates the use of both parameterized general rules in the SDLS of each individual language and then allows for mapping into a common ODLS format. In this way, the divergences are captured, while the ultimate meaning representation is still kept language-neutral. We believe, however, that in the bulk of divergence cases generality is not to be expected, especially in the case of construction divergences. In order to maintain the level of conventionality in translation, an extensive construction lexicon has to be maintained.

5. References

Arnold, D., S. Krauwer, L. des Tombe and L. Sadler. 1988. Relaxed compositionality in machine translation. Proceedings of 2nd TMIMT. Pittsburgh. June.

Dorr, B. 1992. Classification of machine translation divergences and a proposed solution. *Computational Linguistics.*

Fillmore, C., P. Kay and M.C. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of *let alone. Language*, **64**: 501-38.

Meyer, I., B. Onyshkevych and L. Carlson. 1990. Lexicographic Principles and Design for Knowledge-Based Machine Translation. CMU-CMT Technical Report 90-118.

Nirenburg, S. and L. Levin. 1991. Syntax-driven and ontology-driven lexical semantics. Proceedings of the 1991 SIGLEX Workshop, Berkeley.

Nirenburg, S. and I. Nirenburg. 1988. A Framework for Lexical Selection in Natural Language Generation. Proceedings of COLING-88, Budapest, Hungary.

Talmy, L. 1985. Lexicalization patterns: semantic structure in lexical forms. In: T. Shopen (ed.), **Language Typology and Syntactic Description III:** Grammatical Categories and the Lexicon. Cambridge University Press. 57-149.