[Proceedings of the *Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Colgate University, Hamilton, New York, August 14-16, 1985]

LINGUISTICS AND NATURAL LANGUAGE PROCESSING

Victor Raskin
Purdue University

Introduction

The paper addresses the issue of cooperation between linguistics and natural language processing (NLP), in general, and between linguistics and machine translation (MT), in particular. It focuses on just one direction of such cooperation, namely applications of linguistics to NLP, virtually ignoring for now any possible applications of NLP to linguistics, which can range from providing computer-based research tools and aids to linguistics to implementing formal linguistic theories and verifying linguistic models.

Section 1 deals with the question why linguistics must be applied to NLP and what the consequences of ignoring it are. Section 2 provides a counterpoint by discussing how linguistics should not be applied to NLP and, by contrast and inference, how it should be. Section 3 narrows the discussion down to one promising approach to NLP, the sublanguage deal, and the interesting ways, in which linguistics can be utilized within a limited sublanguage. Section 4 is devoted specifically - but necessarily briefly - to the things linguistics can contribute to MT.

Due to the deliberately self-contained design of the paper, only the essential references are listed.

Section 1. Why Should Linguistics Be applied to NLP?

Linguistics is the discipline which is supposed to know about the organization of text. The primary goal of linguistics, according to an enlightened view, is to study the mental mechanisms underlying language. Since their direct observation is impossible, linguistics is trying to match, or model, the native speaker's language competence by observing the indirect consequences of his/her speech output and by discovering and presenting formally the rules governing this output.

What the native speaker is competent about as far as language is concerned boils down to matching sounds and meanings. However, this is done not on a one-to-one basis but rather with the help of a heavily structured medium, consisting of quite a few interrelated levels of interrelated elements. These levels include phonetics and phonology, morphophonology and morphology, syntax, semantics and pragmatics, and text linguistics/discourse analysis.

At each level, linguistics tries to discover and/or postulate the basic units and rules of their functioning. Contemporary linguistics does things formally, which means utilizing one or more - and frequently all - of the various manifestations and/or interpretations of **linguistic formality** listed in (1).

(1) (i) using mathematical notation
(ii) relying entirely on the forms of the words and word combinations in linguistic analysis
(iii) adhering to the mechanical-symbol-manipulation-device paradigm

(1iii) is the strongest and most serious commitment to formality, having far-reaching consequences, free from concern for the cosmetic factors involved in (1i), and less constrained in its heuristics than (1ii) . What the **mechanical-symbol-manipulaticn-device (MSMD)** approach amounts to is all of the factors listed in (2).

(2) (i) collecting all the relevant information without any rigorous method and denying any possibility of formal heuristics
(ii) summarizing the findings as a formal mathematical object, most frequently as a set of rules forming a calculus or a formal grammar
(iii) applying the formal object from (ii) to describe, or generate, a set of linguistic entities sharing a property or properties
(iv) assuming (practically without ever trying to prove that experimentally) that the native speaker has a natural intuitive ability to distinguish each linguistic entity having the property (or properties) in question from any entity lacking it (or them)
(v) claiming (or more realistically, trying to perfect the formal object in (ii) to achieve a situation, such) that the set of entities described or generated in (iii) contains all the entities, to which the native speaker assigns the property (or properties) in question, and nothing but such entities

Chomsky's transformational grammar is, of course, the best known example of an MSMD linguistic theory, and the problematic property it is built upon is grammaticality. However, recently, most formal proposals in linguistics, including the anti-Chomskian ones, have followed the MSMD format and aimed at discovering and/or postulating a set of rules.

It might seem, and may have seemed for a while, that the MSMD format brings linguistics tantalizingly close to computer science and that the rules and sets of rules proposed by the former can be directly implemented by the latter for NLP. It will be shown in Section 2 that "it ain't necessarily so." This, however, should not at all lead to the opposite reaction, displayed by quite a few NLP experts and groups, that linguistics is practically totally useless for NLP.

Everybody who has had some practical experience in NLP knows that at a certain point one has to describe the morphology, syntax, and semantics of a natural language. Not only does linguistics possess most, if not all, of the knowledge one would need in this situation, but much of it is already pre-formated and pre-formalized for the NLP person, though never in his/her favorite format or convenience language. The alternatives to tapping this resource are listed in (3).

(3) (i) using published grammatical descriptions, which are often imperfect and always inconvenient to use
(ii) resorting to monolingual dictionaries, which are nothing short of disaster in coverage, methodology, selection, and consistency (bilingual dictionaries are even worse)
(iii) doing introspection, i.e. using one's own (or an associate's) native competence, which invariably leads to the reinvention of the wheel, and quite often, the wheel does not even come off quite round

In many projects, ignoring linguistics and not employing active research linguists or defectors from linguistics, some combination of (3i-iii) is utilized, and a price is paid for that in efficiency and quality.

Typical examples of linguistic wisdom, necessary for NLP and immediately available to a linguist but not easily accessible, though certainly known in principle, to the native speaker, are listed in (4-19) , roughly according to the level of language structure. Almost all of the examples are related to ambiguity, easily the thorniest issue in NLP.

As far as phonetics/phonology is concerned, unless an NLP system

contemplates the acoustic input/output, which is hardly ever the case and not quite realistic at this point, this level does not have any significance for NLP. Its written correlate, **orthography**, does not have much to offer either, though a sophisticated system might list permissible spelling variants, such as in (4i). Another possibility of utilizing the linguistic knowledge at this level would be treating spelling as self-correcting codes and devising a program which would correct a misspelling to the nearest correctly spelled word in the lexicon. However, in most languages and certainly in English, there are two many pairs of words, such as in (4ii), the distance between which is 1. Treating spelling as an error-detecting (but not self-correcting) code is more realistic if it is based on what might be termed **graphotactics**, similarly to its known oral correlate phonotactics. The latter deals with permissible sequences of sounds in a language; the former would deal with permissible sequences of letters (or other graphemes) in the orthography of a language. A simple program based on graphotactics would rule out strings like the ones in (4iii). However, this would be taken care of also by looking up - and not finding - a word in the system's lexicon if unfamiliar words are unlikely to occur.

(4) (i) fulfil : fulfill, antisemitic : anti-semitic: anti-Semitic, stone wall : stonewall
(ii) read : lead: bead, lane : lake : lace, lace : lack, tie : tee, tie : tip
(iii) *rbook, *tfa, *bkate, *stocm, *haa

Morphemes, the meaningful parts of words, are the minimal language entities which have meaning, and they are in fact the lowest level of language structure which concerns NLP directly, simply because NLP is interested in what the processed text means. Morphonology and morphology are the two levels dealing with the morpheme. **Morphonology** knows that some morphemes have different spellings (and pronunciations) but remain identical otherwise - some obvious examples are listed in (5i). **Morphology** contains data and rules on the various exceptions from seemingly obvious rules. Thus, while thousands of English nouns are pluralized by adding -(e)s to their singular form, quite a few are not - see some representative examples in (5ii). On the other hand, a noun having the standard plural form can in fact be in the singular and require the singular form of a verb to agree with it, e.g., is (5iii); then again, a noun may have the plural form and require the plural form of a verb, e.g., are, but still denote a single object (5iv).

The concept of the zero morpheme is not trivial either - in (5v), the lack of an additional form in the first word of each group is as meaningful as the underlined additional morphemes in the other words. The zero morpheme in the three listed cases means 'noun, singular, non-possessive,' 'verb, present, non-3rd person singular,' and 'adjective, positive (non-comparative, non-superlative) degree,' respectively. One also needs to know that the same morpheme in a language can have multiple meanings, each determined by its position and function. Thus, in (5vi), the same English suffix -s means 'verb, present, 3rd person singular,' 'noun, plural, non-possessive,' and the apostrophe has to be counted as a regular character in order to distinguish either of these two forms from the two possessive forms, plural and singular.

(5) (i) capable : capability, serene : serenity, incredible : impolite
(ii) many children, sheep, syllabi, formulae, addenda
(iii) news, linguistics, statistics

(iv) scissors, trousers
(v) boy, boys, boy's, boys' ; walk, walks , walked, walking; white, whiter, whitest
(vi) walks , books; student's , students'

A linguist also knows, without having to figure it out, that there are parts of speech, such as Noun, Verb, Adjective, etc., and that each of them has a typical paradigm of word forms, listed in (5v) for the three parts of speech in question. Somewhere on the border of **morphology and syntax**, another piece of wisdom, potentially of a great interest for NLP, looms large, namely that the same morpheme in English can signify a different part of speech as in (6).

(6) (i) John saw a big stone
(ii) in some Arab countries they stone adulterous women to death
(iii) this is a stone wall

Only a syntactic analysis of each sentence or at least of a part of it - and not a simple morphological characteristic in the lexicon - can determine whether the word in question is a noun, a verb, or an adjective.

In **syntax**, the available wisdom is even more varied, and complex. A few less obvious examples are listed in (7). (7i-iii) are typical cases of syntactic ambiguity, paraphrased as (8i-ii), (9i-ii), and (10i-iii), respectively. (7iv-v) are two sentences which have a different surface structure but the same (or very similar) deep structure. (7vi-vii) are the opposite examples - the surface structure is the same but the deep structures are different; (10iv-v) illustrate the difference. (7viii-x) contain a verb which must be used with maximum one noun phrase (the subject only), minimum two noun phrases (the subject and the direct object), and minimum three noun phrases (the subject, the direct object, and the indirect object), respectively.

(7) (i) flying planes can be dangerous
(ii) old men and women
(iii) time flies
(iv) the dog bit the man
(v) the man was bitten by the dog
(vi) John is eager to please
(vii) John is easy to please
(viii) John snores
(ix) John sees Mary
(x) John reminds Mary of Bill
(8) (i) it is possible that flying planes is dangerous
(ii) it is possible that flying planes are dangerous
(9) (i) old men and old women
(ii) old men and age-unspecified women
(10) (i) one does not notice how much time has passed
(ii) you there, measure the performance of flies with regard to time
(iii) a breed of flies called 'time'
(iv) John pleases somebody
(v) somebody pleases John

In **semantics**, the most important item for NLP is the homonymy of words and ambiguity of sentences. Dealing with the written language, NLP has to be concerned not only with full homonyms (11i), which are spelled and pronounced

the same way and have different and unrelated meanings, but also with homographs (11ii), whose pronunciations are different, and with polysemous words (11iii), whose meanings are different but related.

(11) (i) bear$_1$ 'give birth' : bear$_2$, 'tolerate' : bear$_3$ 'wild animal'
(ii) lead 'be the leader' : lead 'heavy metal'
(iii) bachelor 'unmarried man; academic degree; subservient knight; etc.'

Homonyms, homographs, and polysemous words are the usual source of purely semantic ambiguity (12i), as opposed to the purely syntactic ambiguity in (7i-ii). (7iii), however, was an example of a mixed, syntactico-semantic ambiguity, which is very common, because both the syntactic structure of the phrase and the meanings of the two words are changeable (time is polysemous, and flies homonymous). Semantics is connected with syntax and morphology in other ways as well: thus, the animal meaning of bear, bear$_3$ in (11i) is excluded from consideration for (12i) because it is a noun, while the syntactic structure of the sentence determines the slot as a verb.

(12ii) exhibits a much more sophisticated kind of referential/attributive ambiguity, which tends to be overlooked by non-linguists almost universally and which is important for NLP, for instance, from the point of view of whether a token in the world of the system need to be actualized or not. (13-14) parahrase the ambiguous sentences of (12i-iii), respectively.

(12) (i) she cannot bear children
(ii) John would like to marry a girl his parents would not approve of
(13) (i) she cannot give birth
(ii) she cannot stand children
(14) (i) there exists such a girl that John would like to marry and his parents would not approve of her (referential)
(ii) John would only like to marry such a girl that her parents would not approve of her (attributive)

While almost any sentence can be ambiguous, hardly any is intended as ambiguous in normal discourse. What it means is that disambiguating devices are available to the speaker and hearer. Some of them are in the text itself, others are in the extralinguistic context, and linguistics is supposed to know about both but, in fact, knows much more about the former. (15i) contains a well-known example of a sentence containing a homonymous word, bill, with at least three meanings, namely, invoice, legal, and bird-related, and in (15ii), it is disambiguated with the help of another word, paid, which corroborates only the invoice meaning. Priming procedures in NLP are based on this and similar kind of corroboration.

(15) (i) the bill is large
(ii) the bill is large but does not need to be paid

Two words corroborate, or prime, each other's meanings if they share one or more semantic features, and the concept of semantic feature is central to contemporary semantics. In various ways, it has been incorporated into a number of formal semantic theories and into quite a few NLP lexicons. Thus, bill and paid in (15ii) share the feature of 'money related' or whatever else it might be called.

The processing of a text by the native speaker and, therefore, by the computer as well depends heavily on a number of even more complicated

meaning-related items which are studied by pragmatics, the youngest and least developed area of linguistics. It is known in **pragmatics** that the same sentence can play different roles in discourse (known as the illocutionary forces - see Searle 1969), and pragmatics studies the factors which determine the roles in given situations. (16) can be perceived as a promise, a threat, or a neutral assertion, depending whether the hearer would rather the speaker came home early, would rather the speaker did not come home early, or does not care when the speaker comes home.

(16) I will be back early

(17) contains an example of a sophisticated and little explored role-related ambiguity. The same sentence (17iii) in a dialog can signify agreement and disagreement, depending on whether it is uttered in response to (17i) or (17ii), respectively. The resulting polysemy of no is not obvious to most native speakers.

(17) (i) the weather is not too nice over there
(ii) the weather is nice over there
(iii) no, it isn't

Pragmatics is also interested in situations, in which sentences are not used in their literal meanings, i.e., as implicatures - see Grice (1975). Thus, (18i), which is phrased and structured as a question, is in fact typically used as a polite request. (18ii) may be used sarcastically about an idiot. (18iii), though ostensibly laudatory, may be a sexist put-down.

(18) (i) can you pass me the salt?
(ii) he is a real genius
(iii) she cooks well

The examples in (17) involve two-sentence sequences, (17i) followed by (17iii) or (17ii) followed by (17iii). The structure of such sequences of sentences, of paragraphs, which are supposed to be logically organized sequences of sentences, and of whole texts, which are sequences of paragraphs is the major concern of text linguistics/discourse analysis (with the second term, extremely homonymous in its use, usually emphasizing the structure of dialogues, or conversational strategies). This discipline is somewhat older than linguistic pragmatics but even less definite about its facts or methods. Some of the simplest examples of sentential structures are such sequences as the enumeration in (19i), the temporal sequence in (19ii), and the causal one in (19iii) - the underlined words are the connectors, which provide explicit clues as to the type of the structure.

(19) (i) The English verb paradigm contains four basic forms. First, there is the infinitive form, which doubles up as the non-3rd person, non-singular form of the present. Secondly, there is the 3rd person, singular form of the present. Thirdly, there is the past form, which double up as the past participle form with the regular verbs. Fourthly and finally, there is the gerund form, which doubles up as the present participle form.
(ii) In the morning, I get up at 6. Then I take a shower and have breakfast.
(iii) I cannot fall asleep as easily as other people. Because of that, I try

to avoid drinking strong tea or coffee after 6 p.m.

All of the listed and many similar pieces of linguistic knowledge (4-19) are more or less immediately accessible to a linguist, though some do require more sophistication, e.g., (12ii) and (17-19). All of them are related to ambiguity and, therefore (at least potentially) important for NLP. One serious problem for NLP is that all of these facts cannot be found in any one published source and certainly not in any acceptable form, and the only way to obtain them all when they are needed is to have a linguist around on a permanent basis. Now, the reason the written sources do not exist is not because the linguists keep the knowledge to themselves so as to sell it - and themselves - to the highest bidder but rather because of serious theoretical problems, some of which are inherent only to linguistics while others are shared with the other human studies. It is essential, therefore, for any NLP project with some concern for adequacy and efficiency, to have a linguist on the staff. A much more serious problem for NLP is that having a regular linguist on the staff is not good enough.

Section 2. How Not to Apply Linguistics to NLP

A linguist on the staff of an NLP project should have an immediate and errorless access to all the linguistic facts of the kind listed in Section 1 and of potential or actual importance to the project. Now, much of this information comes packaged as part of a formal grammar, i.e., as a set of rules. The linguist should be smart enough to know that the packages are not ready for use in NLP. In fact, much of the negative attitude to linguistics on the part of NLP researchers is due to the fact that once they obtained such a package by themselves and tried to implement it directly, because it looked formal and even algorithmic enough to be implementable, without the benefit of a weathered linguist's advice.

The weathered linguist differs from a regular linguist and even from a good regular linguist in that he or she knows the rules of correct linguistic application. These consist of general rules of application and specific rules of linguistic application.

Generally, when a source field is applied to the target field, it is essential that the problem to be solved comes entirely from the latter, while the concepts and terms, ideas and methods, and the research design as a whole may be borrowed from the former. If the problem comes from the source field, the application is not likely to yield any insight into the target area, nor will it be of any value to the source field either because, in most cases, it does not need any additional proof that a certain method works. Thus, it is clear that statistical methods can be applied to anything that can be counted. It may be perfectly possible to determine, with a great degree of reliability, which country in the world leads in the number of Jewish-Gypsy couples, which have two or more children and an annual income over $21,999, and by how much, but unless this answers a real question in demography, ethnography, and/or economics, the research will be a (statistical exercise) in futility.

Similarly, linguistics can, for instance, analyze any sentence syntactically and do it pretty well. It would be rather unwise to hope to get a handle to poetry and to claim that linguistics is being applied to poetics if all one did was to analyze every sentence of a poem syntactically. On the other hand, if poetics comes up with a real question concerning the role of syntactic structure in achieving a certain kind of rhythm or effect, the same syntactic methods can be used fruitfully, and a correct application of linguistics to poetics will be taking place.

In other words, as far as linguistics as the source field and NLP as the target field are concerned, no purely linguistic problem should be imposed on NLP and substituted for a real NLP need. No linguistic method should be used or linguistic description attempted to be implemented unless this is necessary for the realization of the project. Now, all of this is different if the project is, in fact, about a research model in linguistics and the computer implementation aims entirely and deliberately at verifying a linguistic model or description or checking the linguistic formalism. This is the only situation, in which a straightforward implementation of, for instance, Chomsky's transformational grammar would make any sense. It is quite possible that some useful results may be obtained in the course of this kind of work for regular, non-linguistic-research-model NLP, but these gains are likely to be indirect and almost tangential. The linguistic-research models will be ignored here for the rest of the paper.

For a real-life, non-linguistic-research project in NLP, aiming at a working system, the typical dilemma is that a good linguistic description is needed but without the forbidding-looking, cumbersome, and inaccessible packaging it typically comes with. The weathered linguist should unwrap the package for his/her NLP colleagues, separate the gems of wisdom from the wrapping which, at best, answers some purely linguistic needs, and let the group utilize the "real thing." In order not to perform that kind of operation from scratch and on an ad hoc basis every time it is needed, the NLP-related linguist should be able to rely on an applied linguistic theory, specially adapted for NLP. This is exactly what computational linguistics should be about but for the most part is not.

An applied linguistic theory for NLP should contain formulae of transition from linguistic theories and models to models and descriptions practically digestible for NLP. It should be able to distinguish between elements of language substance and the purely linguistic representation of them, not necessarily of much use for NLP. It should be able to take into consideration the state-of-the-art methods and tools of implementation in NLP and the convenience of implementing various kinds of linguistic information with their help. In other words, such a theory should have the beneficial effect of repackaging the linguistic goodies NLP wants in the way, which is the most convenient for NLP to use.

As an example, Postal's classic and sophisticated treatment of the English verb remind (1971) will be compared with what NLP is likely to need to know about it. Focusing on just one meaning of the verb as used in (20i) and deliberately excluding the meaning in (20ii) from consideration, Postal comes up with a number of sharp even if at times controversial observations about the verb, briefly summarized in (21). He then proceeds to propose a transformational treatment for the sentences containing the verb in the likeness meaning, again briefly summarized in (22). The sentences triggering and/or resulting from the transformational process are listed in (23).

(20) (i) Harry reminds me of Fred Astaire
(ii) Lucille reminded me of a party I was supposed to attend

(21) (i) the verb remind must be used with exactly 3 NP's in one particular syntactic structure, viz., NP_1 Verb NP_2 of NP_3
(ii) remind differs syntactically from the other very few English verbs which can be used in this structure
(iii) remind is unique in that no two of its three NP's can be coreferential
(iv) sentences with remind in the likeness meaning are typically paraphrased as, for (20i), (23i)

(22) (i) the standard transformational generative processes are assumed to have generated a structure like that of (23i)
(ii) a transformation, called 'the psych movement,' interchanges the subject and object of the higher sentence in the structure, yielding a structure like (23ii)
(iii) a transformation, called 'the remind formation,' changes (23ii) into (20i)

(23)(i) I perceive that Harry is like Fred Astaire
(ii) *Harry strike me like Fred Astaire

(24) Harry is like Fred Astaire

Typically for the best transformational work and very elegantly, the choice of transformations is determined primarily by the unique feature of remind (21ii). It is demonstrated that each of the three non-coreferences involved is not unique and is, in fact, derived from one of the transformations applied to generate (20i). One non-coreference follows from presenting the sentence as a two-clause structure with (24) as the lower clause, with similarly non-coreferential NP's. Another follows from the psych formation, motivated independently on other English material. And the last and most problematic non-coreference is shown to follow from the remind formation, which is, of course, postulated specially for the task and thus not independently motivated as a whole but, in its components, related to various other independently motivated rules.

The point of the description is that the verb remind is derived transformationally and therefore does not exist as a surface verb. That was supposed to prove that the claims of interpretive semantics concerning deep structure and lexical insertion were false.

NLP will ignore both the theoretical point of the previous paragraph and the entire contents of the one before it. What NLP, or the applied theory catering to it, should extract from the entire description and discussion can be briefly summarized as (25).

(25) (i) remind has (at least) two distinct meanings illustrated in (20)
(ii) = (21i)
(iii) = (21iv), elaborated as (iv)
(iv) NP_1 reminds NP_2 of NP_3 = NP_2 perceive(s) that NP_1 is (are) like NP_3 = = it strikes NP_2 that NP_1 is (are) like NP_3

The difference between what linguistics wants to know about the English verb remind and what NLP must know about it has a deep theoretical foundation. Linguistics and NLP have different goals, some of which are presented schematically - and necessarily simplistically - on the chart in (26).

(26)	Linguistics Wants:	NLP Needs:
(i)	to know all there is to know about the complex structure mediating the pairings of sounds (spellings) and meanings in natural language	to use the shortest and most reliable way from the spellings to the meanings in the text(s) being processed
(ii)	to structure linguistic meaning and relate it to context	to understand the text and make all the necessary inferences
(iii)	to distinguish the various levels of linguistic structure, each with its own elements and relations	to use all the linguistic information which is needed for processing the text(s) without any concern for its source

(iv)	to draw a boundary between linguistic and encyclopedic information to delimit the extent of linguistic competence and, therefore, the limits of the discipline	to use encyclopedic information on par with linguistic information, if necessary for processing the text(s)
(v)	to present its findings formally, preferably as a set of rules in an axiomatic theory	to implement the available information in a practically accessible and convenient way

The situation is complicated by the fact that, in most cases, linguistics cannot offer the definite, complete, and conclusive knowledge of the facts. Thus, in spite of the enormous and concentrated effort in transformational grammar since the early 1960's, no complete transformational grammar of English or any other natural language has been written - a fact, which often surprises and disgusts NLP researchers but should not. If, for instance, linguistics had fulfilled (26ii), the processes of understanding in NLP could follow the resulting structure of meaning. In reality, NLP can only incorporate the abundant but fragmentary semantic findings.

To ignore linguistics in this situation may be simpler than to use it. It is also extremely wasteful and self-defeating. To apply linguistics fruitfully and correctly, one has to be both a well-trained and weathered linguist and an accomplished NLP-er. More realistically, a working tandem of a linguist, knowledgeable about NLP and willing to shed some of his/her theoretical arrogance, and a person in NLP, enlightened enough about linguistics to be respectful but firm enough to be demanding, would be a good solution to the dilemma presented in this and the previous sections. (If everything else fails, they can at least have an interesting discussion along the lines of Nirenburg (1985).) One particular form of linguistic application is briefly discussed in the next section.

Section 3. Sublanguage

One significant difference between linguistics and NLP is that while the former is concerned with language in general, the latter deals with a(n often extremely) limited part of it. In fact, the difference is much less pronounced when one realizes that, on the one hand, in practice, a linguist also deals with the descriptions of very limited fragments or manifestations of language while on the other hand, serious NLP research always aims at significant generalizations about the whole problem. The difference is more in the emphasis on what is typically done in either field. If the linguist had to describe a particular language or its part every time he or she wanted to publish something, the problems would be at least partially very similar to the practical headaches and hard choices faced by NLP when working on a parser and a lexicon. If, on the other hand, an NLP researcher could get away with simply theorizing about the problem, he or she would probably move much closer to linguistics - in fact, those scholars who do, do.

Typically, an NLP project deals with a limited sublanguage of natural language, such as the language of a narrow area of science or technology. By doing that, NLP puts linguistics even further on the spot because to be useful, it would have to shed its most important, though for the most part unconscious idealization, namely that one native speaker's competence is identical to any other's. It is true that there are areas in linguistics, such as dialectology, sociolinguistics, and - recently and most unsurely of itself - linguistic pragmatics, which do not subscribe to the idealization. However, the bulk of linguistics ignores the obvious fact that, in a certain empirical sense, the

Chinese, English, Spanish, Hindi, Swahili, Russian, etc., languages do not exist. What exists instead in reality is the 700 million or so Mandarin Chinese dialects, 400 million or so English and Spanish idiolects, etc. What follows is that the rules formulated for a language may not be true of many of its dialects and idiolects; the lexicon of the language is not utilized in its entirety by any of its native speakers; the syntactic inventory available in the language is used only partially in any dialect.

It is obvious, nevertheless, that the national language exists in some less empirical and more abstract way in spite of all that. However, theoretically this situation is not easy to resolve, and linguistics has largely ignored it. Raskin (1971) seems to remain the only monograph on the subject, and even that effort was geared towards a computational aim. In more practical terms, some recent efforts in NLP are characterized by a growing realization of the predominantly if not exclusively sublanguage orientation NLP (see, for instance, Kittredge and Lehrberger 1982) and of the need to take advantage of the situation without shooting oneself in the foot.

What happens practically when dealing with texts from a limited sublanguage is listed in part in (27).

(27) (i) the lexicon of the sublanguage is limited to just a few hundred words, which is a mere fraction of 500,000 or so words in the maximum dictionary of a full-fledged multi-register national language
(ii) the amount of homonymy and polysemy is reduced drastically because many meanings of potentially troublesome words go beyond the sub-language in question
(iii) the amount of extralinguistic knowledge about the world described by the sublanguage is many orders of magnitude smaller than the global knowledge of the world
(iv) the inventory of syntactic constructions, available in the language, is used only in small part in the sublanguage

Thus, none of the words in (28i) is likely to occur in textbooks or research papers on NLP. The words in (28ii) will lose all of their numerous computer-unrelated meanings. The piece of common-sense knowledge in (28iii) will never be used. The syntactic structure in (28iv) is unlikely to occur in any text of the sublanguage.

(28) (i) beige, whore, carburetor, serendipity, jejeune
(ii) operate, data, user, insert, memory
(iii) a person, considered good-looking, is likely to attract sexually other persons, primarily of the opposite sex
(iv) that bad - what a shame - oh, all right, what can one do?

There are two undesirable extremes in dealing with sublanguages. The first one is to ignore their limitedness and deal with each as if it were the entire language. It would seem that nobody would be likely to do that, especially given the fact, mentioned at the end of the previous section, that linguistics typically does not furnish complete descriptions of the entire languages. It is surprising, therefore, to discover many traces of the (largely unconscious) language-as-a-whole approach, manifesting itself usually as the descriptions of phenomena, which cannot occur.

The other extreme is much more widespread because it is tempting and - in the short run - efficient. Following it, one tends to describe only what is

there in the texts being processed, in a highly ad hoc fashion, which makes it impossible to extrapolate the description beyond the sublanguage and which makes the system extremely vulnerable in case of the occurrence of any slightly non-standard text or even individual sentence within the same sublanguage. Thus, it would be foolish to process the word xerox in a sublanguage entirely on the basis of its being the only word in the lexicon beginning with an x. More plausibly, it would be near-sighted, in the computer sublanguage of English, to take advantage not only of the fact that the verb operate has lost all of its computer-unrelated meanings, such as the surgery meaning, but also of the fact that its only direct object in the sublanguage is computer. A non-ad-hoc solution would be to define it in this meaning as having something like machine as its direct object and to make computer the only child of machine in the sublanguage. Then, in case of an extrapolation, it may be easier to add children to machine than to redefine the verb. In general, an extrapolation is much simpler to bring about with the help of a mere addition than by restructuring the description.

A wise approach to sublanguage in NLP requires, therefore, not only that information elicited from linguistics be mapped onto NLP needs but also that it be reduced in size, as it were, to ensure an economical but non-ad-hoc description of the linguistic material.

It appears that theoretical research on sublanguage has also the most to offer to MT as a specific problem in NLP.

Section 4. Linguistics and MT

Linguistics should be able to contribute to MT in two ways. First, within its general contribution to NLP as outlined above, since MT is primarily NLP, albeit with its own specific problems not necessarily shared by other areas of NLP. Secondly, MT should profit from an application of linguistics to a general theory of translation, no matter whether human or automatic. Only the latter aspect will be briefly commented upon in this section.

Unfortunately, linguistics has had very little positive to say about translation. In fact, in the early literature on MT in the 1950's, those who claimed to be speaking for theoretical linguistics (or for the philosophy of language - see Quine 1960) argued against the feasibility of any MT and deplored any practical endeavors in this direction as impermissible short cuts, having nothing to do with the way language was. While they may have been right most of the time then, the unhelpful, standoffish attitude, resulting virtually in no attempt to look at the problem of translation from a serious linguistic perspective, was surprising. One explanation of that phenomenon could be the very limited constrains on linguistics at that time and the antisemantic attitude of the then dominant structural linguistics.

A much broader view of linguistics at present and the wealth of semantic and pragmatic wisdom accumulated in the last two decades or so should have changed the situation, and it is true that these days, one notices more literature on translation appearing. However, most of the effort comes not from linguists but rather from philosophers and philosophically minded literary scholars (especially, from the more formal schools of literary criticism) and practitioners. Much of the literature remains anecdotal, and the concerns expressed are usually of a stylistic and/or aesthetic nature.

It is true that translation is not a linguistic problem - it is extraneous to the discipline. However, to the extent that translation involves the use of one or more natural languages, what linguistics knows both about language in general and about the involved language(s) cannot be ignored. Similarly to the reasoning in Section 2, the only chance for linguistics to contribute to

translation is via an applied linguistic theory catering to the needs of the field.

What is the main problem of translation? It can be presented as the ability to determine whether some two texts, each in a different language, are translations of each other. In order to be translations of each other, the texts should probably satisfy the following linguistic conditions (29):

(29) Two texts in different languages are translations of each other if they have the same:
(i) meaning
(ii) illocutionary force and perlocutionary effect
(iii) inferences

Obviously, (29i-iii) are interrelated, while focusing on general and specific facets of meaning. The term 'perlocutionary' (see Austin 1962, and Searle 1969) is used here as an extension beyond linguistics of the notion 'illocutionary,' i.e., the role of an utterance in discourse. Perlocution covers the extralinguistic effect of the text on the hearer and his/her resulting actions, moods, attitudes, etc. Perlocution is determined also by the additional factors in (30) but those go definitely beyond linguistics and into stylistics, rhetoric, and composition, respectively (to each of which linguistics can also be profitably applied, though again on a carefully limited scale).

(30) Two texts in different languages are translations of each other if they have the same:
(i) stylistic status (e.g., scholarly style)
(ii) rhetorical effect (e.g., persuasive)
(iii) aesthetic effect (e.g., well-written)

(It is interesting to note that the conditions in (29-30) are equally applicable to two texts in the same language, i.e., paraphrases of each other.)

Given the goal of linguistics to match the native speaker's competence, the applied linguistic theory of translation should aim at matching the bilingual native speaker's translation competence, which, of course, can only be done practically by observing and studying their performance. These observations will yield interesting results. It will become clear immediately that there is a many-to-many correspondence between texts in one language and their translations in the other. The differences between any two alternative translations will be primarily due to syntactical and semantical variations. The word-for-word translation is ruled out by morphological differences as well, and the more sophisticated morpheme-for-morpheme approach will not work out either. In the decreasing degree of triviality, (31) lists various deviations from the morpheme-for-morpheme approach in translation, and (32) illustrates them with English/Russian examples.

(31) (i) there is no one-to-one correspondence between morphological forms in two different languages
(ii) syntactic structures cannot generally be copied from one language to another
(iii) due to differences in semantic articulation, the same word may be translated differently in two sentences
(iv) an element of meaning may have to be lost in translation
(v) an element of meaning may have to be added in translation

(vi) significant changes in translation may be due to the necessity to control the 'given-new,' or 'topic-focus' information

(vii) a significant rephrasing may be necessary for illocutionary reasons

(viii) additional information of a sophisticated pragmatic kind, e.g., the different systems of honorifics, i.e., forms of address depending on the speaker/hearer's (relative) status, may determine the outcome of translation

(32) (i) walk (V) = xodit', xozu, xodiš', xodim, xodite, xodjat;
walk (N) = progulka, progulki, (o) progulke, progulku, progulkoj;
he walked, had walked, was walking, had been walking = on guljal

(ii) the train being late, he missed the meeting = poskol'ku poezd opozdal, on propustil zasedanie /because the train was late.../

(iii) they are romantically involved = oni neravnodusny drug k drugu /they are not indifferent to each other/
Russia is heavily involved in Nicaragua = Rossija sil'no zamešana v delax Nikaragua

(iv) I washed my hair = ja vymyl golovu /I washed head/

(v) the sky was blue = nebo bylo goluboe /light blue/
are these shoes black or blue? = éti tufli černye ili sinie? /dark blue/

(vi) a man came into the room = v komnatu vošel čelovek /into room came man/
the man came into the room = čelovek vošel v komnatu /man came into room/

(vii) can you pass me the salt? = bud'te dobry, peredajte sol' /be (so) kind, pass salt/

(viii) "I love you," Count Ebuchev whispered to Princess Poblyadushkina = = "Ja ljublju vas /polite you/", prošeptal graf Ebucev princesse Pobljaduskinoj
"I love you," said Evdokim the shepherd to Agrafene the dairy maid = "Ja tebja /familiar you/ ljublju", skazal pastux Evdokim dojarke Agrafene

The best contribution linguistics can make to translation, besides merely alerting translators to the factors in (31) and the other similar ones, is by providing, via the applied theory, the format for bilingual descriptions and by filling this format with information for each pair of languages.

One would think that linguistic universals should also play an important role in translation by facilitating it. It is true that translating into a non-human language, i.e., an artificial or space alien language, is likely to be much harder. However, most universals are of a highly formal nature and a very limited practical use (e.g., the universal specifying that each natural language uses entities of three levels, sound, word, and sentence)

The transition from a linguistic contribution to translation in general to a linguistic contribution to MT involves primarily the selection function. While many translations of the same text are possible, they are usually weighted on the scale from optimal to barely acceptable. The selection function assigning the weights is determined by the factors in (30) and other factors concerning, for instance, the special purpose of the text, i.e., to have a poetic effect. In MT, due to the limited nature of most projects, the selection function may be often allowed to stay strictly within the basic requirements in (29). It has been demonstrated in earlier work (Raskin 1971, 1974) that in addition to that, in limited sublanguages, some of the factors in (31) do not apply or at least not to the same extent. Thus, as far as (31ii) is concerned,

all the permissible syntactical transformations of the same sentence - and in a limited sublanguage, the inventory is greatly reduced - can be treated as identical, and therefore, any variant will do, at least as long as (31vi) is not affected. (31iii) may be dropped altogether thanks to the limited lexicon. (31vii-viii) are extremely unlikely to play any significant role, either.

References

Austin, John 1962. How to Do Things With Words. New York - London: Oxford University Press.

Grice, H. Paul 1975. "Logic and conversation." In: P. Cole and J. L. Morgan (eds.), Syntax and Semantics, Vol. 3. Speech Acts. New York: Academic Press, pp. 53-59.

Kittredge, Richard and John Lehrberger (eds.) 1982. Sublanguage: Studies of Language in Restricted Semantic Domains. Berlin - New York: de Gruyter.

Nirenburg, Sergei 1985. "Linguistics and artificial intelligence." In: P. C. Bjarkman and V. Raskin (eds.), The Real-World Linguist: Linguistic Applications in the 1980's. Norwood, N.J.: Ablex (in print).

Quine, Willard V. 0. 1960. Word and Object. Cambridge, MA: M.I.T. Press.

Raskin, Victor (V.) 1971. K teorii jazykovyx podsistem /Towards a Theory of Sublanguages/. Moscow: Moscow University Press.

Raskin, Victor 1974. "A restricted sublanguage approach to high quality translation." American Journal of Computational Linguistics 11:3, Microfiche 9.

Searle, John R. 1969. Speech Acts. Cambridge: Cambridge University Press.