Learning to Translate :
A Psycholinguistic Approach to the
Induction of Grammars and Transfer Functions

Patrick Juola

CU-CS-767-95 April 1995

University of Colorado at Boulder

Technical Report CU-CS-767-95
Department of Computer Science
Campus Box 430
University of Colorado
Boulder, Colorado 80309

Learning to Translate : A Psycholinguistic Approach to the Induction of Grammars and Transfer Functions

Patrick Juola

April 1995

Abstract

Human language is one of the most intricate and complex systems with which scientists have tried to work. Many projects have foundered on the complexity of natural language and the difficulties of describing, in a principled way, its regularities and its idiosyncracies. Mathematical formalisms capable of describing its generality have proven infeasible to learn.

Machine translation, translating automatically from one (natural) language to another, is in an even deeper hole, because of the difficulties of dealing with two sets of idiosyncracies simultaneously. In theory, all the information that a translator needs can be obtained from a set of already translated text—but it has proven very difficult and time consuming to write programs that are capable of working with this sort of information, and the most successful systems use naive and linguistically implausible formulations that are almost impossible to understand, modify, or use.

Linguistic typologists and psycholinguists have identified many constraints on the form and processing of human languages. By incorporating these constraints into a language learning system, it is possible to build a system that learns to translate (infers functions and grammars for machine translation) from an aligned bilingual corpus of sentences using understandable, symbolic linguistic principles and representations. This work focuses on one particular constraint, the Marker Hypothesis, which is shown to be powerful, understandable, and computationally accessible.

This hypothesis has been incorporated into a family of systems that infer such transfer functions using standard multivariate optimization techniques. These systems have been tested on a variety of language pairs and corpora, demonstrating the language and corpus independence of this approach. Furthermore, the design principles are in theory independent of any particular inference technique or grammatical representation and reflect only the constraints of the Marker Hypothesis and similar psycholinguistic principles.

Because of the symbolic nature of this approach, the transfer functions learned are easy for non-mathematicians to use and modify. It is equally easy to apply other sources of linguistic information to help speed and direct the learning task. This can make the task of developing machine translation systems much simpler and represents a significant improvement over the current state of the art.

Part I
Introduction

1 Problem statement

This thesis presents the results of several experiments in developing a system that learns to translate between natural languages by a linguistic examination of translated text. As presented here, the METLA system infers a grammar and symbolic transfer functions from an aligned bilingual corpus of sentences. More accurately, the system infers a set of parameters which collectively describe a grammar and transfer functions. The inference, in turn, is deliberately designed to incorporate the use of psycholinguistically plausible linguistic universals to guide and simplify the inference task in the interests both of speed and understandability of the final results.

More specifically, this work focuses on the computational implications of one universal in particular. The Marker Hypothesis[27] has long been known to aid humans in the learning of miniature artificial languages; the conjecture here is that it will also aid computers in the learning and processing of natural languages. Furthermore, it is sufficiently obvious, in the sense of not requiring tremendous theoretical machinery, that it can be easily applied to the learning task.

To this end, this work describes two particular instantiations (METLA-1 and METLA-2) of this translation task. Both are similar in their general structure, but use different linguistic formalisms to describe the source language and the parsing task. METLA-1 uses a computational formalism based on the application of the Marker Hypothesis to standard phrase-structure grammars, while METLA-2 applies the Marker Hypothesis to a more lexical formalism called Categorial Grammar[80] (CG). Both formalisms have been extended and modified to incorporate the restrictions and constraints suggested by the Marker Hypothesis. The METLA systems also infer a context-sensitive bilingual dictionary and a set of restructuring relations to define the actual translation process itself. These inferred functions, in turn, are developed and written in a form that, as will be shown, is easy to apply to novel sentences, simple to understand, and very straightforward for non-mathematicians to modify or improve.

2 Motivation

Translation and its associated difficulties have been around as long as separate human languages. For thousands of years, one of the marks of the educated has been the ability to read and understand foreign languages. In the modern community, the ability to communicate in foreign languages is essential to scientific research, diplomatic communications, and commercial development. Unfortunately, the ability to read and write a foreign language is a difficult task, typically requiring years of study to acquire even a moderate level of skill. The ability to speak multiple languages fluently and to translate between them is rare and in high demand. Human translation is a multibillion dollar industry in the United States alone, despite being expensive, time-consuming, and labor intensive. For this reason, the idea of programming a computer to translate documents between languages has been a popular and highly-studied one from the beginning of the computer era in the early 1950's.

To lay some groundwork about the difficulty of translation, consider simply trying to define the notion of "translation." A simplistic definition would treat the translation process as the process of trying to express the meaning of a sentence in one language (here and after, the "source" language) in another (the "target" language). The task of a translator (human or machine) is then to select words and constructionss that mean the same thing in the target language as the words and constructionss given in the source. Unfortunately, this leads immediately into extremely treacherous ground. It has been proposed and argued[57, 61] that the notion of equivalence of meaning is improper. The cognitive aspects of this approach are unpromising, and from an engineering standpoint, we find ourselves immediately faced with the problem of complete world knowledge.

As an example, consider the collection of meanings encompassed by the English verb *"to wear"* (as in "to wear [clothing, jewelry, etc.]")." In Japanese, these meanings are spread out across five verbs, and selection among these is based on a number of semantic and lexical features.[1] It is clearly a difficult and time-consuming task to encode these features into a form that a computer can use; in fact, it is not straightforward just to devise a principled way to predict which verb a given item will take. However, the information needed to learn these distinctions is available in the huge amount of already-translated text available from various sources worldwide.

The now-mainstream Example-Based Translation paradigm (e.g. [53, 62]) relies on exactly this sort of information to produce translations with a reduced but still significant amount of human engineering. By encoding a tremendous database of sentences or fragments together with their translations, a system can select the most appropriate (for example, closest to the source sentence) fragment or fragments and combine them into a novel sentence. However, in addition to the engineering work necessary to collect and store the fragments, the notion of "most appropriate" has proven to be slippery and requires a tremendous amount of world knowledge to be stored as well. In [62], for instance, the engineers incorporated a commercial thesaurus in their attempt to classify nouns into various categories. It would be a major improvement to develop a system capable of identifying significant features of a bilingual corpus and structuring a transfer function around them. Systems which have been built to perform this task, however ([7, 40]) tend to be linguistically implausible and for this reason difficult to understand and hard to modify and maintain.

By incorporating psycholinguistic principles into the inference task, it is possible for a system to learn the significant structures and features of a language using a representation that is linguistically plausible and understandable. Furthermore, it should be possible to use this representation to solve other natural language problems, such as machine translation. This work focuses on the computational implications of one major psycholinguistic result, called the Marker Hypothesis. As presented here, the METLA systems can derive transfer functions that result in reasonably accurate translations. The transfer functions themselves are, as will be shown, simple to understand in linguistic instead of mathematical terms, psycholinguistically plausible, and easy to update and modify.

It should be noted that this work does not, formally speaking, constitute an "experiment" to test the Marker Hypothesis, except in a vague and exploratory sense. In particular, there is no control group for the performance of the METLA systems to be measured against. The

[1] A shirt and similar objects worn on the torso takes a different verb than a sock or shorts, worn below the waist or a hat, worn on the head. Jewelry takes a fourth verb, while eyeglasses (and blankets) appear to have a word unto themselves. This description, is, of course, highly oversimplified.

reasons for this are several. First, as discussed in part IV, the Marker Hypothesis has been proposed as a universal of human language. It is therefore not necessarily sensible to discuss the idea of a language without (suitably general) "marker constructs." Second, the problem of generalized language induction is known (see section 9) to be impossible without some source of additional information, and any amount of positive performance at all represents some sort of performance increase. Third, a major part of the motivation for the METLA system is not the numerical performance increase, but the demonstrable linguistic plausibility of the inferred representations. This research should be viewed as a proof-of-principle, that the Marker Hypothesis can be used as an adjunct to general NLP processing, rather than as a direct measurement of the sorts of improvement that can be expected.

3 Overview

Parts 2 and 3 describe some of the relevant linguistic and computational background and related work, including definitions of most of the terms and formalisms used in the remainder of the thesis. Parts 4 and 5 discuss psycholinguistic universals, paying particular attention to the Marker Hypothesis and its mathematical and computational implications. Part 6 discusses, in a broad sense, the nature of a system that can be built to use the Marker Hypothesis to perform induction of grammars and transfer functions. Parts 7 and 8 describe two particular implementations of that system, with the results of several experiments in a variety of languages and corpora and their evaluation. The extensibility and future development of the system is discussed in parts 9 and 10. Part 11, in turn, summarizes the entire research and concludes with a note on appropriate future directions for natural language processing research.

Part II
Linguistic Formalisms

4 Introduction

One of the most important questions for linguistics can be phrased quite simply as "What's the difference between an ungrammatical sentence and a grammatical one?" For example, why is the sentence "Anthony bought a bunch of flowers for her" acceptable, while "*[2]Anthony a bunch bought of flowers for her" not? This question can be answered at a number of levels. On one level, the word "bought" is not allowed to be followed by the word "of." A more useful and linguistically cogent explanation is that "a bunch of flowers" is a unit (a prepositional phrase) and cannot be broken across a verb. This second explanation is more useful because it relates the ungrammaticality above to many other phenomena in many other languages, instead of being an isolated fact about two words. In general terms, languages display regularities in their ordering and selection of words which can be examined and described by linguists in formal, language and vocabulary independent, terms.

5 Phrase-structure grammars

Modern formal syntax theory began with Chomsky[12]. Prior to this, grammatical sentences were generally analyzed by psychologists in terms of Markov chains. (Linguists' descriptions roughly resembled those in language textbooks, which are so informal as to be computationally useless.) Determiners, for example, are generally followed either by adjectives or by nouns. Adjectives, in turn, are followed either by other adjectives or by nouns, and nouns are often followed by verbs. In general, the next event in a Markov process can be predicted by observing the event or events that came immediately before it, without regard to lengthier contexts. Although mathematically tractable, this approach is clearly inadequate to handle the sort of dependencies commonly found in natural languages.

For an easy demonstration of this, consider the following set of sentences :

The box is red.
The boxes on the table are red.
The box on the table beside the dresser is red.
The boxes on the table which my sister built beside the dresser are red.

In all cases, the number of the verb (is/are) must agree with the number of the subject noun (box/boxes), despite an indeterminate and possibly unlimited number of words, enough to exceed any possible (finite) context for a Markov chain. Chomsky demonstrated by a similar argument that human language could not be predicted by Markov chains and elaborated[12] a much more powerful formalism called *phrase-structure grammars* that captures long-distance dependency that is not permitted to Markov chains. A formal definition of these grammars and their properties can be found in [30, 31]; the discussion here will be limited to the direct results applicable to natural language processing and specifically to context-free grammars.

[2] Here and elsewhere, '*' denotes an ungrammatical or otherwise unacceptable utterance.

Under this theory, a language can be formally defined as a set of finite-length strings (sentences) over a finite alphabet. Sentences which are in the set are grammatical, sentences which are not members of that set are ungrammatical. Associated with each language is a *grammar* generating it. Formally, a grammar may be defined as a 4-tuple (T,V,R, S), where T is a set of *terminal symbols* that may appear in the language, V is a set of *nonterminal symbols* or *variables* that do not appear in the language but are used to generate the language, R is a set of production rules of the form $\phi \rightarrow \psi$ (where $\phi \in (V \cup T)^+ - T^+$ and $\psi \in (V \cup T)^*$), and S is a designated element of V called the *start symbol.* Any rule of the form $\phi \rightarrow \psi$ can be used to produce a new string y from an old string x by direct substitution of the string ψ for ϕ at one location where it occurs in x. The language generated by the grammar G, termed $L(G)$, is the set of all strings that contain only elements of T that can be generated in one or more steps from the unit-length string S. For an example of this, consider the following grammar (Γ). Upper case letters are non-terminals, lower-case letters are terminals, and S is the designated start state. The rules of the grammar are

S → AbrAcAdAbrA
A → a
A → wxpm

This grammar generates [L(Γ) is] 32 different strings of the form 'abracadabra', 'wxpmbracadabra', 'abrwxpmcadabra', 'abracadwxpmbra', 'abrwxpmcwxpmdabra', and so forth. A sample derivation might run as follows

S → AbrAcAdAbrA → abrAcAdAbrA → abrAcwxpmdAbrA
→ abrAcwxpmdabrA → abracwxpmdabrA → abracwxpmdabra

The differences among various classes of languages are expressed as restrictions upon the productions of a grammar. As discussed above, a Markov chain is a process where the next event can be described as a function of the immediately preceding event(s) without regard to a longer context. Under Chomsky's formalism, such processes can be described by so-called regular grammar, where every production is of the form $A \rightarrow \alpha B$ or $A \rightarrow \alpha$, where A,$B \in V$ and $\alpha \in T^*$. The lefthand side of the rule contains only a single character, which mush be a nonterminal, and the replacement string can have at most one nonterminal in it, which must be the final symbol in the replacement. (Γ, above, is not a regular grammar.) Successive relaxations of this very powerful restriction result in, respectively, context-free grammars, context-sensitive grammars, and unrestricted grammars. These categories together constitute the Chomsky hierarchy. A language L is said to be regular (context-free, etc.) if and only if there is a grammar G that generates L and G is regular (context-free, etc). This should not be interpreted to mean that if L is a regular language, every grammar that generates L must be regular—since this is a strictly increasing hierarchy, every regular language is also a context-free language (and so forth), and so might also be generated by a more powerful grammar.

6 Context-free grammars

Context-free grammars, the second rung of the Chomsky hierarchy, represent a relaxation of the restrictions on regular grammars. Specifically, they require that only one nonterminal be replaced at a time (productions must be of the form $A \rightarrow \beta$ where $A \in V$ and $\beta \in (V \cup T)^*$), but can be replaced with any string. (Γ, above, is a context-free grammar.) This allows them to express long-distance dependencies by the production of two nonterminal symbols at once (for example, a sentence can be either a plural-noun/plural-verb combination or a singular-noun/singular-verb combination), but makes it much more difficult, given a particular word, to determine which other words are important to its context.

From a linguistic perspective, phrase-structure grammars (and particularly context-free grammars) are a mixed blessing. No human language is known to be simple enough to be completely described as a context-free language, but many projects ([78, 4] among many others) have been built using subsets of natural language modeled as context-free grammars (with or without additional information). Some languages, in fact, are known to have constructs that are provably not context-free. Swiss-German and the phenomenon known as "cross-serial dependency"[56] illustrates this. On the other hand, CFGs provide an excellent framework for the discussion and representation of high-level linguistic phenomena and have formed the basis for most theories of syntax since they were introduced. Furthermore, the induction of context-free grammars compatible with a given corpus has been a major focus of research in computer science, as will be summarized in section 9 under the heading "Grammar acquisition."

7 Categorial grammar (CG)

Despite the successes of the phrase-structure grammar formalism, there are other approaches for the formal descriptions of syntax. Some researchers, for example, may feel uncomfortable with the implicit entities described by nonterminal symbols. Others point to the formal and practical difficulties of learning a grammar (as children are said to do) from only examples of sentences. Others consider the entire notion of formal grammar to be an inappropriate mechanism to describe syntactic structures. For this reason, other formalisms have been developed. One of the more useful, from a computational perspective, is called *Categorial Grammar.*

First developed by Bar-Hillel[3] and used by many later researchers (e.g. [80]), CG can be seen as a method of describing syntactic regularities purely on the basis of properties of the individual words in a language. The basic units of analysis are entities (nouns, typically), and propositions (sentences), which can carry truth-values about entities. Other words are described in functional terms—for example, a transitive verb is simply a function which converts two noun phrases into a sentence. The basic structure of English derives partially from the fact that the two noun phrases expected by a transitive verb are on opposite sides of the verb itself, while in Urdu, the two noun phrases expected by a transitive verb are both to the left of the verb.

Wood[79] gives an example of this process reproduced here as figure 1. The notation $x/(y)[z]$ describes the lexical (sub)categorization associated with a word. In particular, (y) means that this word expects to be immediately preceded by a word or phrase of category y, while $[z]$ describes similar expectations for the following word or phrase. Either or both of these sorts of expectations may be present; if all are met, the resulting conjunction is of category x. As hinted above, the basic categories for this particular analysis are 's' (for sentence) and 'n' (used here for "everything else")—the exact choice of atomic categories is a hotly debated topic among CG researchers.

In brief, words such as 'man', 'John', and 'Paul', which represent entities, are categorized as 'n'. Words like adjectives ('poor') and determiners are categorized as 'n/[n]', functions mapping a noun to another, more complex noun. Only one parse is available for phrases such as "a poor man," reflecting the necessity for every modifier to have an object to modify. Verbs,

```
John      knew      that    Paul      was        a     poor     man
  n     s/(n)[n]    n/[s]     n     s/(n)[n]   n/[n]   n/[n]      n
  n     s/(n)[n]    n/[s]     n     s/(n)[n]   n/[n]          n
  n     s/(n)[n]    n/[s]     n     s/(n)[n]              n
  n     s/(n)[n]    n/[s]                       s
  n     s/(n)[n]                            n
                                  s
```

Figure 1: Sample grammatical analysis in CG

as have already been discussed, are mappings from one or more nouns into sentences, while the complementizer 'that' can convert a sentence into an entity in its own right for further discourse. From a formal standpoint, CG is clearly of equivalent power to standard CFGs—every lexical item can be easily converted into a CFG production, and conversely, every production in a CFG in Greibach normal form (see pg. 35) can be described as a CG lexical subcategorization for its initial nonterminal.

Categorial grammar defines an easily-computable but linguistically viable approach for describing syntactic structures. Because of its strong focus on the lexicon, it is relatively simple for a computer to identify a potential set of categorial expectations or to define novel words in terms of other words with similar categorization. These properties may make CG an ideal formalism for the investigation of corpus-based linguistic phenomena.

Part III
Computational background

8 Introduction

"Natural Language Processing," or its synonym "Computational Linguistics," are both terms describing that subfield of computer science that tries to operate on and with natural (human) languages. Grammar acquisition, here defined as the process of learning a formal grammar compatible with a known set of sentences, is clearly a part of NLP, as is machine translation. By combining results from these approaches in conjunction with known properties of human languages, a system can be built to automatically infer translation functions from a database of translated sentences. This chapter focuses on some of the major developments in these two areas that are relevant to the development of the METLA translation system.

9 Grammar acquisition

A significant theoretical limitation on language learning (defined by computer scientists as discovering arbitrary patterns in strings of symbols) was discovered by Gold[26] in 1967. He defined a class of languages as "learnable in the limit" if and only if there is an effective procedure (i.e. a computer program) that will allow a learner to correctly identify a particular element of the class and the learner will not change her identification with more information. Gold found that no language class that contained all languages of finite cardinality and at least one language of infinite cardinality was learnable from positive examples only. In particular, Gold's results show that no part of the Chomsky hierarchy (see section 5) is learnable from positive examples alone. As human languages are, in theory, more general than the formal languages of the Chomsky hierarchy, and as they also, in theory, could contain any finite language (and are presumed infinite in extent by linguists), this theorem should apply as well to a computer trying to identify a human language.

Of course, as will be discussed in part IV, this mathematical result does not stop children from learning their native language – and therefore, computer acquisition of human languages may still be possible if the problem is sufficiently restructured. A typical research project will avoid the hangman's noose of Gold's theorem in any of several ways. The most common approach is simply to provide negative data, either in the form of a judge who can state whether or not a proposed sentence is a member of the language, or simply to provide explicitly labeled negative data. Examples of this approach date from 1964 ([70, 71]) and continue on to the present day ([18]). Unfortunately, this is not a good model of the process by which humans learn language. Humans neither receive nor act upon explicit negative instances when learning language ([46, 65]). A more subtle version of this technique involves the use of implicit negative data (as in [58]), where a positive instance of one thing is presumed to be weak evidence against all other concepts, but this requires the system to be embedded in a larger cognitive framework and is difficult to apply to the strict problem of grammatical induction.

Another common approach is to identify a subset of context-free languages and demonstrate a learning algorithm that works on that subset. This can sometimes be justified as an assumption about the structure of the languages examined. [1, 5, 15, 44] are all examples of this approach. Unfortunately, the subsets are usually chosen for their mathematical properties and are not particularly useful for translating natural languages. For example, [44] identified an algorithm that will identify "hierarchical" grammars—grammars not containing recursive rules. This restriction, unfortunately, immediately rules out constructions of the form "John said that [Mary told him that [...]]" where the construction following "that" is a full sentence—a very common construction in natural languages.

One interesting subset that has been extensively studied is the so-called "structured samples" as defined by Crespi-Reghizzi [15] and further examined in [2, 55, 42]. These are structures where all subcomponents marked by brackets, or, alternately, sentences which have undergone immediate constituent analysis. In other words, every rule of the form $A \rightarrow x$ has been replaced by a rule $A \rightarrow (x)$ where "(" and ")" are symbols that do not appear in the original grammar. This produces sentences like "(((ab)(bc))(cd))", where "abbc", and "cd" are the subcomponents of the larger sentence "abbccd," and "abbc" is itself further subdivided into the components "ab" and "bc." Grammatical induction of structure samples is equivalent to accepting unlabeled parse trees instead of mere sentences, and the learning task is merely to label the parse trees correctly. This is much easier than the task of learning the parse trees from markers, but is compatible with certain known psycholinguistic principles, specifically the Marker Hypothesis as will be defined in section 14.1. Another interesting variation is are the so-called "pattern languages" studied by Angluin[1]. These languages are a subset of the regular languages, produced by the use of composition and the Kleene plus[3] operator, that can be shown to be learnable from only positive data. The guarantee of at least one occurrence of every pattern in the language allows a marker-like algorithm to deduce the underlying pattern in the words of the language. However, both of these languages are obviously highly restricted and neither would work as a good model for human language structure and acquisition.

A third approach, typified by Berwick[4] and Miikkulainen[49], involves the identification and/or marking of thematic categories, and using this information to generate parsing and generation algorithms. This method is unfortunately inappropriate for self-organizing translation systems since the identification of thematic roles is itself a semantic problem no easier than translation and requires human expertise or an extremely restricted grammar.

The final point to be addressed is whether the results from grammatical induction are completely relevant to this project. Pat Langley[41] has pointed out that it seems more plausible that a child learns to map meaning structures from the syntax she hears, rather than just learning that some strings are grammatical and others are not. If the meaning of the string can be extracted semi-automatically (rather than by hand-encoding thematic roles), then it may be possible to determine enough information to perform translation. The translation results from [7] (described below) are obtained using Markov chains and no notion of grammar. In addition, even systems which purport to use grammatical information to perform their translations should be able to work in the presence of parsing errors. Humans frequently utter malformed or fragmented sentences, particularly in conversations where the missing parts can be identified in context. The translation of a malformed sentence should be a (possibly malformed) sentence in the target language, not a simple statement that the source sentence had an error. Grammatical information should be an aid to translation, not a set of handcuffs on the translation system. So the inability of a system to completely and accurately identify the generating grammar of a CFL may not be a devastating handicap for a translation system.

[3] a symbol meaning "one or more occurrences of"; see [31] for formal definition

10 Automatic translation

The major engineering bottleneck to machine translation, in general, is the development of the knowledge base. Even using human translators, the development of a cross-linguistic dictionary is typically a large fraction of the cost of a translation problem (and something that human translators will typically not do, instead requiring their clients to supply it with the source document). The development of transfer rules and the appropriate circumstances to apply them is typically a major project, requiring as much labor as the development of any similarly-sized expert system. It should therefore come as no surprise that a major focus of research has been on ways to reduce the amount of human thought that goes into translation, the so-called "knowledge acquisition bottleneck."[62]

There are three "traditional" approaches to machine translation (see [36, 67, 54] for a more detailed survey of machine translation techniques), characterized by increasing linguistic sophistication as opposed to purely syntactic transformations (reflecting the increased capability of linguistic, AI, and compiler technologies available). The simplest, termed "direct translation," is best represented by SYSTRAN, one of the more widely used machine translation tools in the United States. Systems built using this approach tend to use simple word-for-word or phrase-for-phrase translations, with only limited ability to rearrange syntactic constructions or select appropriate words from sets of synonyms, generally built as special cases into the translation engine. Most modern systems use the "transfer" approach (for an example, see [23, 32], one of the best-known success stories of this approach), based loosely upon the idea of producing a data structure representing the syntax and semantics of the source text, then transforming this structure into a new one representing the same semantics but the syntax of the target language. The construction of this transfer function, of course, represents a major investment of time and effort. The final approach, called the "interlingua" or "pivot" approach, separates the syntax of the source and target languages still further by parsing the source sentences into a semantic representation completely divorced from the source text, then retelling the sentence in a target language using standard text generation techniques. (See [49] for a recent example of this technique.)

All these approaches require significant human expertise to be programmed into them, whether as special cases in the FORTRAN code or as a full representation (a la Schank [64]) of the semantics of the utterance. In addition, although the more sophisticated approaches may appear to require less time (since they can capture the so-called rules of grammar more directly), they require that the inevitable exceptions to every rule of grammar be coded as well. Like any large rule-based system, the interactions among transfer rules are not always obvious and amenable to correction.

11 Example-based translation

Nagao[53] outlined what would become a major force in machine translation with his proposal of what would become known as "example-based translation." Rather then generating explicit rules, translation systems would store a huge collection of translation examples that would provide coverage in context for the input. For a typical input sentence, the system would parse it into its components, examine the database for examples which were semantically similar to the components of the input sentence, then construct the translated sentence by putting together the translations of the input components according to the grammatical rules of the target language.

A simple example of this technique was developed by Sumita et al. [75, 74] and called EBMT (for Example-Based Machine Translation). This system translated only phrases of the form "$noun_1$ NO $noun_2$," where NO is the general partitive adposition in Japanese. In most contexts, it is similar to the English word "of," but can also be translated as "in" (as in "the conference in Kyoto"), "'s" ("the teacher's pencil"), "by," possessive pronouns, and several other potential translations depending upon the nouns surrounding it. Sumita et al. incorporated a commercial thesaurus into their system and imposed a notion of semantic distance upon the classes of potential nouns the system understood. Using 2250 pairs of Japanese phrases and their English translation, the system would select the semantically closest pair to the desired source sentence and translate the source phrase into a structure similar to the matched English phrase.

This method has several distinct advantages over more typical approaches. It is much better at selecting the "correct" translation from a set of synonyms or near-synonyms. It is typically more robust than a fully rules-based translation system and can use (human) translation expertise well. As a further advantage, it can assign a confidence or reliability factor based on the nearness of the semantic matches. On the other hand, it still requires highly developed skill to develop a translation database, and still more skill to develop a suitable parser, generator, and thesaurus for calculating semantic similarity. There is no easy way for a computer to recognize two words as being semantically similar or distinct, which creates a new knowledge acquisition bottleneck.

A similar corpus-based approach was developed by Brown et al.[7]. Using a huge bilingual corpus, they attempted to solve the translation problem as a mapping between Markov chains, asserting that every sentence is a possible translation of every other sentence, then calculating the most probable translation from the statistics of the corpus, defined as the product of the probabilities of each word in the source language producing a particular word or set of words in the target language.

The limitations of such a word-by-word translation approach are many and varied. There is no idea of grammar, only a simple Markov chain. There is no context-sensitivity, and no notion of selecting the most appropriate translation from a set of near-synonyms. This method, however, appears to require relatively little human expertise or time to produce its translations.

A similar approach has been used by Koncar[40], who used a large neural network to infer a translation method between a large set of English sentences and their Serbo-Croat translations. The advantages and limitations of artificial neural networks are well-known; of particular significance to this work is that, again, they are very weak in their ability to handle grammatical constructions, and it is very difficult to analyze the inferred functions in any sort of linguistically plausible or understandable fashion.

A final example of this sort of data mining can be seen in the work of Jones et al.[72, 35] on the automatic extraction of translation functions by alignment. In many regards, this is the most linguistically sophisticated and plausible data-mining system produced to date. It preserves, for example, the notion of compositionality in the sense of the final translation being produced by a compounding of (potentially many) translated fragments. However, the fragments themselves are not produced by any sort of a parse scheme, and are instead produced by simple dynamic programming routine rather similar to the UNIX diff(1) program. The fragments identified are not necessarily linguistically useful or well-formed. Finally, the transfer process itself is nearly as opaque as those produced by Koncar's neural network, as there is no easy-to-describe method for performing the recompounding or rearrangement of the translated texts.

12 Weaknesses of example-based translation

The unifying principle behind these approaches is relatively simple. All the information that a translator needs can be found in existing examples of translated text. It is at least worth investigating whether or not this is a reasonable assumption, and what weaknesses this approach brings.

First, for instance, it should be obvious that a corpus-based NLP system will only learn properties present in the training corpus. Similarly, an EBMT system will learn to translate based on the grammar, lexical choices, style, and so forth, of the bilingual corpus used for training. In other words, if the METLA system is trained on a database of English and Spanish newspaper articles, it will learn to translate from a journalistic style and into a similar journalistic style. This should be no more surprising than the fact that it will learn to translate from English into Spanish. This can be an advantage in some cases. For instance, a fully-functional system will be able to learn certain stylistic regularities ("technical papers never use the word 'I'") that may or may not be part of the grammar of the source and target language—or that may not even be obvious or evident to the scholars who develop the source and target dictionary. Similarly, since the system models the training corpus, new corpora which are filled with "grammatical errors" (such as sentence fragments or sentences-that-end-with-a-preposition, very common in conversations) will be translated as accurately as they are translated in the training corpus. The system will not be handicapped by a prescriptive notion of "grammatical."

At the same time, it should be evident that the more closely the system models one style, the less appropriate it will be for other styles. A system trained on this thesis (and a hypothetical translation), for instance, will learn a large fraction of the grammar of academic English, plus any stylistic quirks present in the original or the translation. These quirks may be inappropriate for a later translation of a Hemingway novel, as his style is very different from typical academic prose. Furthermore, the availability of texts, particularly texts of specific styles, may be limited – for example, EEC official documents of one form or another are very easy to get, but are written in a very formal, legalistic style. Bilingual examples of casual, familiar writing (or worse, casual, familiar speech) are much more difficult to obtain.

A similar problem can be found with the representation of world knowledge, which by definition is outside of any linguistic context. The English word "sister" has two translations in Japanese, one meaning 'older sister' and one meaning 'younger sister.' Only world knowledge about the speaker and the speaker's family is capable of resolving the ambiguity in the sentence "My sister would like to live in Los Angeles," and so a provably correct translation of this sentence is impossible for any corpus-based translation system. Similar problems about context length, such as the correct translations of null subjects into languages that do not permit null subjects, can be equally problematic.

Another major problem with corpus linguistics is the question of the adequacy of the texts. It is often surprising how uncommon certain words actually are. For example, [34] identified the words 'bikini,' 'vinegar,' 'rooster,' 'infinity,' and 'well-designed' as having a frequency of approximately one appearance per million words. A million-word corpus may seem like a lot, until one realizes that this only provides about a 70% chance of including a particular word with a one-in-a-million frequency.

Finally, the major problem with most fully example-based translation systems is their linguistic validity, here equated to the ease in which they can be understood and modified in linguistic terms. It is very easy to classify, for instance, Spanish as a language that "permits null subjects" while French is not. Similarly, English is a verb-medial language while Japanese and Urdu are verb-final. These observations, despite their apparent simplicity, are very powerful and can be a useful aid to learning a language or identifying unfamiliar structures and words during translation. Such insights, however, are impossible to systems such as [40] and [7], with their mathematically tractable but linguistically naive formalisms. This is the major problem I hope to address in the current work. By incorporating psycholinguistic principles into the design and functioning of a corpus-based translation system, a set of robust transfer functions can be developed that are nonetheless linguistically cogent and plausible.

Part IV
Linguistic Universals and the Marker Hypothesis

13 Linguistic universals

From a philosophical perspective, if language acquisition is an unsolvable problem, how do humans solve it? And, from a computational perspective, how can computers approximate a solution? One solution that has been proposed for both is that human languages are not an unlimited set and that there are implicit limitations on the variance of human languages. The search for these limitations and restrictions takes up a large fraction of the effort of linguistic typologists, and many of their results can have important computational implications.

Consider the following examples, taken from Talmy[76] and Croft[16]. Many languages, such as English and French, make a grammatical distinction between singular ("that dog barks") and plural ("those dogs bark"). Other languages, such as Japanese, make no such distinction. Still other languages will distinguish singular (one item), dual (two items), and plural (more than that). In some extreme cases, there are languages which code for the numbers one, two, three, and more than three. However, no language is known that codes distinctions between prime numbers and composite numbers. An efficient computer program to acquire human language, therefore, should not spend much time considering hypothetical languages with a prime/composite number distinction. For this reason, an engineer need not waste her time by building knowledge of prime and composite numbers into this hypothetical computer program. By incorporating Talmy and Croft's observations into the design of the computer program, this observed linguistic universal has provided efficiency increases both by restructuring the search space at run time and by reducing the demands on the system engineer.

In many cases, universals can not only reduce search spaces but can actually structure them. In the example above, there are only four possible slots for number agreement, resulting in a maximum of sixteen possible agreement sets. Greenberg[28] has demonstrated a hierarchy of implicational universals among them—any language which distinguishes a higher number also distinguishes lower numbers. For example, a language which has a special number agreement system for dual subjects will also have a special number agreement system for singular subjects. A language with trial (three items) agreement will also have dual and singular agreement. Thus, a computer program can search only four attested agreement patterns to find the appropriate one for a given language.

Another example of these implicational hierarchies and their possible effects on language acquisition can be seen in the Accessibility Hierarchy as presented by Keenan and Comrie[38]. In a cross-linguistic study of the formation of relative clauses, they found that some grammatical structures are much more easily made into relative clauses.

For example, every language studied permitted subjects of sentences to be relativized upon

The man picked up a hat.
I saw the man that [REL picked up his hat] leave.

Most, but not all, language permitted direct objects also to be relativized upon

The man picked up a hat.
I saw the hat that [the man picked up REL] blow away.

Still fewer language permitted oblique objects

The man went to a party with my cousin.
The cousin that [the man went to a party with REL] lives in Chicago.

Even fewer languages permitted genitives

The man's hat blew away.
The man [whose hat blew away] came in late.

And even English, one of the most versatile languages studied, starts to break down with objects of comparison

The dress is less expensive than the hat.
? The hat that [the dress is less expensive than REL] is in fashion.

(and so forth for several other grammatical constructions.) What Keenan and Comrie discovered is that all languages that permit relativization on oblique objects also permit relativization on all other constructs higher on the list. Similarly, every language that permits genitives to be relativized also permits obliques (and things higher). Again, a well-constructed computer program can simply identify which of the very few possible attested patterns the language under study follows and set a few flags appropriately.

Of course, universals phrased in terms of higher-order linguistic structures may not be very useful for computation. It's difficult to glance at a page of text in a completely unfamiliar language and identify which are the relative clauses and what their syntactic structure is.

Another example of this sort of inaccessibility can be seen in Dorr's work[19]. Her PRINCITRAN system performs machine translation using a carefully hand-crafted set of linguistic principles to guide the parsing and translation in accordance with a powerful linguistic theory called Government-Binding theory[14]. One principle that has been incorporated into Dorr's system (Trace Theory) determines whether a noun phrase can be coidentified with an invisible and unexpressed trace element across more than one intervening clause boundary. This is a very powerful restriction, but is almost impossible to identify automatically without tremendous amounts of information about the structure and semantics of the language. A simple rephrasing of this parameter setting can make the difficulties clearer — this parameter determines a noun phrase's reference to the same object as an imaginary word (which is not present in the sentence itself) can be blocked by an imaginary boundary (which is also not present in the sentence itself). To adequately represent this principle requires that the system be able to correctly infer locations for these null elements and to classify them correctly. This ability has been provided to Dorr's system by carefully hand-coding the universals, which creates a new bottleneck of defining and describing languages in terms of a predefined parameter set. Although there are many universals that have been proposed, not all of them can be easily used by a computer program.

Fortunately, there are more general and useful universals that can easily be incorporated into the most general structure of a language acquisition system. For example, Jackendoff's $\bar{\text{X}}$ Theory raises to a universal the observation that every noun phrase contains a noun, every verb phrase a verb, and so forth. This universal can be expressed as a general condition on phrase structure rules that, in general

$$\text{XP} \rightarrow \{\text{COMP}\}\ \text{X}\ \{\text{SPEC}\}$$

where an "XP" (X phrase) goes to an X (the *head*) with some sort of optional complementizer or specifier (an adjective, quantifier, determiner, and so forth). Another cross-linguistic pattern is that any rule that mentions a noun that isn't part of the constructions of a noun phrase (e.g. PP → PREP NP) accepts full noun phrases instead of merely nouns. This also applies to verbs, prepositions, and so forth. These observations have been incorporated by Chomsky and Jackendoff[13, 33] into a theory and notational convention called "$\bar{\text{X}}$ (pronounced 'X-bar') Theory" which can be easily expressed in computational terms as a restriction on the forms of allowable grammars. Charniak[11], for instance, describes a system[9, 8] he built to identify grammars that focuses on the types of non-terminals that could appear as components within another non-terminal. He cites as an example that English pronouns can be difficult to tease apart from the English verbs by standard PCFG techniques. By imposing the $\bar{\text{X}}$-derived constraint that verbs were not allowed to appear in the expansion of a pronoun-phrase, the grammars inferred became much more accurate, understandable, and plausible.

Greenberg[28] discovered another powerful set of linguistic universals. He found that the so-called "basic word order" of simple declarative sentences, although varying from language to language, is relatively constant within a language. For example, English is a Subject-Verb-Object (SVO) language, meaning that the subject of a sentence comes first, then the verb, and finally the object. Japanese, on the other hand, is a Subject-Object-Verb (SOV) language, meaning that the verb tends to be the final element in the sentence. Again, because there are only a limited number of possible word orders (six are theoretically possible, but the OVS order is unattested in any human language), a computer program could easily and efficiently search through the potential set to determine which one a given language follows. Greenberg's results are more powerful, however. In addition to identifying the basic word order typology, he also demonstrated that many other features of language are related to the basic word order. For example, languages with basic word order SOV (or verb-final languages in general, but OSV languages are extremely rare) tend to have *postpositions* instead of prepositions, where the linking word follows the noun that it links instead of preceding it. Languages with the verb first, on the other hand, tend to have prepositions instead of postpositions. Similarly, verb-final languages tend to place the possessing noun before the possessed noun in genitive constructs, while verb-initial languages tend to place the possessor second. English, which is neither a verb-initial nor a verb-final language, does both. Again, once a language has been identified as verb-initial, verb-final, or verb-medial, this universal can provide powerful heuristics to identify the rest of the grammar.

14 The Marker Hypothesis

14.1 Statement

The final universal to be discussed, and the one which forms the psycholinguistic underpinnings of the METLA translation system, is the Marker Hypothesis as developed by Green[27] and others[50, 51]. In its simplest form, this universal states that natural languages are "marked" for grammar at surface level—that there exists in every language a small set of words or morphemes that appear in a very limited set of grammatical contexts and that can be said, in a sense, to signal that context. As an example of this principle, consider a basic sentence in English :

> The Boulder Faculty Assembly announced a list of ten faculty awards at its Thursday meeting, with more awards for excellence in teaching than expected.

In this sentence, taken at random from a Boulder newspaper, two noun phrases began with determiners, two with quantifiers, and one with a possessive pronoun. The set of determiners and possessive pronouns in English is very small (less than fifteen words, depending upon how one counts[4]), and the set of quantifiers is equally recognizable.[5] Similarly, every word in this sentence ending with '-ed' is a past tense verb. The Marker Hypothesis presumes the converse of these observations, e.g. that words which end in '-ed' are very often past tense verbs, and the word 'the' usually heralds the appearance of a noun phrase. Or, more generally, that concepts and structures like these will have similar morphological or structural marking in all languages.

14.2 Psycholinguistic evidence

Proponents of the Marker Hypothesis go further, however, claiming not only that these "marker words" could signal the occurrence of particular contexts, but that they do—that marker words form an important cue to psycholinguistic processing of structure. Experiments with miniature languages have backed up this claim. When human subjects are presented with the task of learning a small artificial language from sentences in the language, they learn more accurately and faster if the artificial language has cues of the sort described above. Green[27] showed this effect in artificial languages with and without specific marker words as attested in Japanese. Morgan et al.[50] demonstrated it in languages with and without phrase-level substitutions, as of pronouns for full noun phrases. Mori and Moeser[51] examined the effect of case marking on the pseudowords of the languages. In these and other experiments, evidence confirming the Marker Hypothesis was always found.

Other evidence for the psychological utility of marker words can be found in typological evidence. The original statement of the Marker Hypothesis was based upon the typological observation that every natural language has such constructs, whether in derivational morphology or separate marker words. Even pidgins and creoles have such constructs. For example, [65] lists examples from a pidgin called Russenorsk. In this language, sentences tend to be very simple strings of words, without grammatical inflection. Even in this language, however, verbs are marked with a special '-om' marker, which presumably helps hearers of this language identify the basic concept expressed in a given utterance (and from that determine the appropriate roles of the other words in the sentence).

Other psycholinguistic evidence for such the Marker Hypothesis can be taken from child language acquisition. Constructs which are easily and readily marked (e.g., regular verbs) tend to be learned early and strongly, and may even override other irregular forms which have been learned by rote memorization. The classic child's sentence "*I goed to the store" is an obvious example of this sort of overgeneralization. The child has learned that events which

[4] e.g., is 'thy' worth putting into a translation system?

[5] Although in theory there are an infinite number of quantifiers, words like '635' or 'heptillion' are rare and easy to process. See [16, p. 98 et seq.] for a discussion of number markedness.

have already happened are described by verbs marked with the '-ed' morpheme. Slobin[66] lists dozens of psycholinguistic principles that may describe how children focus on important bits of the language to learn. Many of these (for example, "pay attention to the ends of words") are direct descriptions of phenomena the Marker Hypothesis would predict.

Finally, there is psychological evidence about not only the universality of marker words and morphemes, but also about their cross-linguistic similarity. Certainly, such concepts as case marking, gender, and tense seem to be found in a large variety of languages. Talmy[76] suggests that, in fact, there are certain cognitive aspects or concepts that are inherently likely to be expressed grammatically (using marker morphemes or structural cues) and others that are universally expressed lexically. For example, many languages have inflections on nouns to express the number. On the other hand, there is no known language where morphemes exist to differentiate nouns referring to red objects from nouns referring to blue ones. Color, then, is not a concept expressed grammatically. The typological evidence of number agreement, that prime/composite is not a useful grammatical distinction, has already been presented. The implication is not only that marker constructs exist, but that the semantic concepts and distinctions that they express tend to be expressed in other languages by other marker constructions.

14.3 CG and the Marker Hypothesis

Categorial Grammar, of course, could be viewed as the ultimate extension of the Marker Hypothesis, as grammatical structures by assumption cannot exist except as the expectations of lexical items. In fact, CG can be viewed as the other extreme of a continuum from more standard grammatical theories (such as [14]), which the Marker Hypothesis can reconcile.

Standard grammatical formalisms tend to focus on the properties of strings of words, rather than the properties of the words themselves. Huge categories of words are lumped together under the heading 'noun,' for example, and when further categorization is necessary, then features (such as 'count noun' or 'proper noun') are introduced, rather than examining the words themselves. Many features of ambiguity, such as prepositional phrase attachment or garden-path sentences, are simply ignored.

On the other hand, Categorial Grammar and similar formalisms may place too much emphasis on the lexical properties of the individual words. Implicit in the statement that every word has a set of categorial expectations is the possibility that every word may have a slightly different set of categorial expectations, and that the regularities of grammar are simply an accidental conjunction of fortuitous agreement among the categories. This possibility, however, seems not to hold – the number of words with outlandish expectations (the marker words) is very small, stable, and the words themselves are usually frequent.

The major strength of the Marker Hypothesis is that it provides a natural distinction between function words and content words and describes the differences in their respective purposes. It is possible to describe this distinction fairly easily in terms of Categorial Grammars. For instance, (in English) there must be some sort of categorial difference between bare (count) nouns such as 'car' and determined count noun phrases such as "the car". Construction Grammar, as developed by Fillmore[22], describes this in terms of a property called 'maximality', which simply states whether or not a noun phrase can be used to build other constructions. Categorial grammar, on the other hand, would state that the lexical categorization of 'the' includes (np/[n]) — or in other words, that 'the', functionally, takes a bare noun and produces a noun phrase. Adjectives, however, which do not change the maximality of their respective nouns, are categorized as (n/[n]).[6]

In general, this distinction can be extended into a general classification of words. Content words, under CG, appear to fall into one of the following three categories :

- Nouns (entities), which are basal elements of the grammar,
- Verbs, which are functions from something into a sentence (s), and
- other things (adjectives, &c.), which provide semantic but not syntactic information.[7] In CG terms, these are words which convert a structure of one category into a structure of the same category, i.e. n/[n], s/[s], et cetera.

Function and/or marker words, then, are simply words which serve to transform categories (such as np/[n]) or conjoin categories (such as 'and' np/(np)[np]). This observation provides a simple explanation of numerous properties in terms of an easily-formalizable theory. For a simple example, consider the case system in German. Using a naive view of case, the difference between an uncased noun and a cased noun is expressed in the choice of the determiner. The determiner, then, serves as a word to hold and mark the case of the following noun. In functional terms, the determiner can be viewed as a function from uncased nouns to nouns of a particular case, which in turn is subsumed into the subcategorization framework of the verb. A more sophisticated view of case (such as Fillmore's[21]) produces largely the same results – under this framework, the verb assigns case to each of its syntactic constituents and the determiner provides a place for this case to be marked, again providing, functionally, a mapping from uncased nouns to cased ones. Similar arguments could provide functional explanations for quantifiers, prepositions, conjunctions, and most other marker words. A simple generalization to the notion of words, allowing morphemes to have lexical expectations and subcategorizations, would provide an equally powerful explanation of marker morphemes.

It can be concluded that the Marker Hypothesis can also be used to bolster the psycholinguistic plausibility of grammatical formalisms, and that these grammatical formalisms, can, in turn, be used as an explanation of certain grammatical properties associated with the Marker Hypothesis. Assuming, then, that the Marker Hypothesis is an accurate description of a useful property of natural languages, it is reasonable to use this property in an attempt to build a system that will naturally acquire the grammar of the source and target languages of interest to a machine translation system. The next chapter describes a computational formalism to do exactly that, based on the notion of acquisition of translation functions from large corpora as described above.

[6] I am grateful to Dr. Laura Michaelis for her lucid explanation of how Construction Grammar approaches this.

[7] This is almost the definition of a content word under the Marker Hypothesis.

Part V

Computational implications of the Marker Hypothesis

15 General remarks

What useful properties would marker words[8] have? As described above, they may signal grammatical structure *if* the information can be properly teased out. Smith and Witten[68] used a related hypothesis about "function words" to infer a grammar describing a large corpus. Observing that function words tend to be among the most common words in all languages, they gathered the most frequent 1% of all word types in a large document and used a series of strong statistical filters and n-gram models to infer a grammar of English. In their words, "the result is a relatively compact grammar that is guaranteed to cover every sentence in the source text that was used to form it." In addition, they found that the inferred grammar was plausible under current syntactic theories, unlike many large-corpora projects. These results were obtained despite a relatively inaccurate and implausible definition of function word. For example, one of the top 1% of the words in *Moby Dick*, unsurprisingly, is the word "whale," A more accurate definition, perhaps including length and morphological complexity criteria, would presumably result in more accurate and general inductions.

Another advantage of the Marker Hypothesis, particularly with regard to translation, is the way it isolates content words, which tend to have few translations. Although the many words to many words problem in translation is difficult, most of the difficulty originates not in the translation of words like "computer" or "kidnapping" but in words like "of" or "the." Context-dependencies are typically defined in terms of the syntactic nature of the surroundings, i.e. in terms of the marker words, and can therefore be solved with a more complex theory of marker word translation rather than a more complex theory of translation in general. Finally, Talmy's theories of grammaticalization[76] suggest that the structures indicated by marker words, although they may vary in their structure between languages, will indicate the same general properties and can therefore be incorporated into the target marker words rather than requiring major vocabulary shifts in the target language.

16 Marker-normal form

How, then can the Marker Hypothesis be formally incorporated into a computational theory of language in a way that allows it to be easily used? As described in section 14.1, the crucial property for this work is the existence of identifiable classes of marker words. Specifically, the formalism and system as described below assumes first that the languages of interest can be approximated by a context-free grammar, and second, that these languages can be naturally described by CFGs in *marker-normal form*, as defined below.

The computational background to this project can be summed up in the following mathematical result :

[8] Or morphemes. The current work only focuses on marker *words*, but future developments may include morphological analysis from large corpora as a part of marker identification[29].

Theorem 1 *To every CFG* Γ *there corresponds an equivalent grammar in marker-normal form, where every production is of one of the following forms :*

$$A \rightarrow \epsilon$$

$$A \rightarrow a$$

$$A \rightarrow A_0 a_1 A_1 a_2 A_2 \cdots$$

$$A \rightarrow a_1 A_1 a_2 A_2 \cdots$$

(As usual, upper-case letters are nonterminal symbols, lower-case letters are terminal symbols, and ϵ *is the null string of zero length.) All right-hand sides of productions are either a single terminal symbol, or are an alternating sequence of terminals and nonterminals.*

Proof: Greibach's theorem[30, 31] states that for every CFG, there exists an equivalent grammar in which all productions are of the form $A \rightarrow aBCDEF\cdots$ (so-called *Greibach-normal form*). For a grammar in this form, all right-hand sides consist of exactly one terminal symbol, followed by zero or more nonterminal symbols. For any grammar of interest, begin by finding an equivalent Greibach-normal form grammar Γ for it. This will then be transformed into an equivalent marker-normal form grammar.

Replace every production $A \rightarrow a\beta$, where β is a string of two or more nonterminal symbols, with two productions involving a new nonterminal : $A \rightarrow aX$ and $X \rightarrow \beta$. At this point, all productions involving the original nonterminals of Γ are in the required form for marker-normal form.

Now, consider a variable B that appears in the right-hand side of a rule $X \rightarrow \beta$. If B is the left-hand side of several production rules, create multiple production rules for the nonterminal X with the right-hand side of each production rule for B. Repeat this process with the Cartesian product of all original nonterminal symbols of Γ. At the end of this process, every rule of the original grammar with multiple nonterminals has been replaced with a rule of the form $A \rightarrow aX$, with a single (marked) nonterminal variable. As the right-hand side of X did not contain any of the new nonterminals, every nonterminal in has been replaced by a marked nonterminal, and so the right side of every X-production is also completely marked.

To convert this entirely to marker-normal form may require the addition of another nonterminal between two terminals in the right-hand side of a production. Simply add the rule $\Omega \rightarrow \epsilon$, for a novel nonterminal Ω, and replace all such right-hand sides γ with $\Omega\gamma$ creating the initial nonterminal as required. This grammar is clearly in marker-normal form and also clearly equivalent to the Greibach-normal form grammar from which it was derived. *(q.e.d)*

Theorem 1 has an immediate corollary to reduce the necessary size of the production rules, at the expense of the number of such rules :

Corollary 1 *To every CFG there corresponds an equivalent grammar form, where every production is of one of the following forms :*

$$A \rightarrow \epsilon$$

$$A \rightarrow a$$

$$A \rightarrow A_0 a_1 A_1$$

$$A \rightarrow A_0 a_1 A_1 a_2 A_2$$

Proof: Exercise 4.11 of [30, p. 66] states that every context-free language can be generated by a grammar of the form

$$A \rightarrow a$$

$$A \rightarrow aB$$

$$A \rightarrow aBC$$

The construction of Theorem 1, when applied to the above grammar, produces a grammar of the desired form. If necessary, an ϵ-generating non-terminal symbol can be prepended to any production rules that do not already begin with one. *(q.e.d)*

17 An example

An example of this transformation may be useful. Consider this grammar, which generates sentences that consist of one or more sequences of balanced parentheses around a '+' character.

$$\begin{array}{ll} S \rightarrow (E)S & S \rightarrow (E) \\ E \rightarrow (E) & E \rightarrow + \end{array}$$

Converting this grammar to Greibach-normal form produces :

$$\begin{array}{ll} S \rightarrow (EPS & S \rightarrow (EP \\ E \rightarrow (EP & E \rightarrow + \\ P \rightarrow) & \end{array}$$

All productions are already in marker-normal form except for the rules $S \rightarrow (EPS$, $S \rightarrow (EP$, and $E \rightarrow (EP$. Create two new nonterminals X and Y such that, respectively, $X \rightarrow EPS$ and $Y \rightarrow EP$.

Making this replacement, the (new) grammar becomes :

$$\begin{array}{lll} S \rightarrow (X & S \rightarrow (Y & \\ E \rightarrow (Y & E \rightarrow + & \\ P \rightarrow) & X \rightarrow EPS & Y \rightarrow EP \end{array}$$

As S and E both have two possible productions, X has four first-level expansions and Y two. Performing these yields

$$\begin{array}{llll} S \rightarrow (X & S \rightarrow (Y & E \rightarrow (Y & E \rightarrow + \\ P \rightarrow) & & & \\ X \rightarrow (Y)(X & X \rightarrow (Y)(Y & X \rightarrow +)(X & X \rightarrow +)(Y \\ Y \rightarrow +) & Y \rightarrow (Y) & & \end{array}$$

and a similar substitution should be performed for Y. Finally, adding the rule $\Omega \rightarrow \epsilon$ and padding with Ω as necessary yields a final version in marker-normal form :

$$\begin{array}{llll} S \rightarrow (X & S \rightarrow (Y & E \rightarrow (Y & E \rightarrow + \\ P \rightarrow) & \Omega \rightarrow \epsilon & & \\ X \rightarrow (Y)\Omega(X & X \rightarrow (Y)\Omega(Y & X \rightarrow +\Omega)\Omega(X & X \rightarrow +\Omega)\Omega(Y \\ Y \rightarrow +\Omega) & Y \rightarrow (Y) & & \end{array}$$

This construction clearly results in much larger and potentially less-coherent grammars than the more standard Chomsky- and Greibach- normal forms. However, the Marker Hypothesis implies that explicitly marked grammars such as these are more psychologically plausible and thus that these grammars are likely to be more natural and understandable for human languages. In particular, natural language should tend to have relatively simple descriptions in which the set of terminal symbols that appear alone in productions is distinct from the set of terminal symbols that appear in a marking context; in other words, that the set of marker words is distinct and identifiable. The existence of marker-normal form provides a framework for attempting to solve natural language problems by focusing on the marker words. In addition, the symbolic, plausible, and understandable nature of these grammars makes it easier to incorporate other principles (such as $\bar{\text{X}}$ Theory) into the grammar.

Part VI
METLA : Principles and Design

18 The problem

Given a bilingual text, it should be possible to extract enough information from the text to translate novel sentences. As discussed in section 11, this is the basis of the example-based translation approach. Given that the source and target text are both written in natural language, with its well-known and well-studied properties, it seems only natural to use those properties to guide and improve the extraction process. The METLA system is a family of experimental prototypes to investigate some computational applications of psycholinguistic principles and constraints to the problem of automatic learning of machine translation.

Specifically, the hypothesis presented here is that the application of these principles will result in several significant improvements over the current state of the art. For example, the transfer functions produced by METLA are more linguistically cogent and plausible than those produced by purely statistical methods (e.g. [40]). At the same time, the amount of (human) work necessary to produce a system for a new language pair is be greatly reduced over that necessary for more typical MT systems, and it is easy to incorporate new linguistic principles and phenomena as they are discovered. The final, and potentially the greatest, new strength is that the system can operate with partial linguistic descriptions. For example, it should be easy to modify a working system to incorporate new vocabulary items. Similarly, the system can start from an incomplete set of linguistic parameters (e.g. language X is known to be a subject-object-verb language, with postpositions and mandatory case marking on adjectives), and use them to speed and/or direct the inference process for a more accurate and easily found set of transfer functions.

19 Design considerations

The METLA system infers a grammar and symbolic transfer functions from an aligned bilingual corpus of sentences. More accurately, the system infers a set of parameters which collectively describe a grammar and transfer functions. These parameters, in turn, are derived from and express psycholinguistic theories and constraints. These parameters include :

- A context-free grammar or equivalently strong formalism describing the source language
- A context-dependent bilingual dictionary describing the relationships among lexical types in the two languages
- A set of permutation relations describing the necessary syntactic reconstruction to convert sentences in the source language into their translations in the target language

The system begins with a random set of parameters describing a skeletal grammar and transfer functions. Over many (potentially millions or billions of) passes through the training corpus, the parameters are tuned to reduce the differences between the translated source sentences (as translated by the current transfer functions) and the correct translation as given in the training corpus. The final set of tuned parameters can then be tested for generalization and/or used in a standalone translation system.

Once the system has been tuned to an appropriate set of parameters (or during the tuning phase as part of performance measurement), the parameters are used in a generalized translation function as follows. The parse formalism is applied to an appropriately sized unit of text, typically a sentence, to produce a parse tree. Each leaf of the tree is translated by looking up the appropriate translation in the bilingual dictionary, and then leaves are successively permuted and concatenated until the entire tree has been concatenated into the desired target sentence.

The exact nature of the components in this general description is undefined. For example, the parsing could be done in a variety of ways and the tuning and inference task can be performed by any of a dozen algorithms. This work focuses on the design and development of two instantiations of the METLA framework (METLA-1 and METLA-2) which are functionally identical but differ in the exact makeup of their subsystems and parameter sets. In particular, the grammatical formalism of METLA-1 and METLA-2 differ, resulting in different parsing schema and different methods for storing grammatical information.

20 Components of METLA

The components outlined above will be discussed in detail, along with formalisms that allow them to described in numerical terms and as such optimized for the greatest match between the translated source sentences and the desired target sentences.

20.1 Source grammar

The first step in the translation process, obviously, is to come up with a description of the source sentence(s) in some form amenable to further processing. By assumption and design, this should be something psycholinguistically plausible while still being easily inferrible. In practical terms, this means a context-free grammar or an equivalently strong formalism, at a minimum. The exact nature of this formalism and the parsing algorithm used comprises the main difference between the two METLA systems (METLA-1 and METLA-2) developed and discussed in this work, and for this reason will be examined in detail in the chapters on the individual systems (parts VII and VIII). In brief, METLA-1 uses a greedy, top-down parser that is a direct extension of the marker-normal form formalism, while METLA-2 uses a more general and powerful formalism based on Categorial Grammar.

In either system, the final parameters constitute a formal description of the syntactic properties of the lexical items that can be used to parse novel sentences in preparation for the restructuring and translation phases of the process.

20.2 Syntactic reconstruction

As discussed in section 13, languages differ fundamentally in the syntactic structures that they use to represent similar semantic concepts. A single language, though, usually displays a relative regularity in its structures. For example, English is an SVO language, while Japanese is an SOV language. This cryptic notation refers to the basic word ordering[28] of a simple declarative sentence with a transitive verb, such as "John broke the glass." Syntactically speaking, then, such English sentences are to be composed of a (subject) NP, a verb and an (object) NP, in that order. A similar Japanese sentence would be composed of two NPs and a verb, in that order. Assuming that the grammar developed above can successfully parse and identify the two NPs and the VP from an English sentence, each of these components can be translated as a unit and their translations conjoined in a different order to form the Japanese translation. Numbering the components from left to right, the Japanese is produced by appending the first, third, and second components (after translation). This operation is immediately recognizable as a simple permutation.

A similar permutation could be carried out at every point of application of every grammatical rule in the source grammar. By repeating this translate-permute-concatenate operation recursively, any sentence in the source language can be restructured into a corresponding target structure. To complete the translation process, then, it remains only that the recursion have a base case, i.e. that the individual lexical items be translatable.

20.3 Context-dependent bilingual dictionary

Neglecting the difficulties inherent in world-knowledge and pro-drop languages (see below), every bit of semantic information expressed in the source sentence must be present in the target sentence. The difficulty arises from the possibility of a different and ambiguous encoding in either or both languages. For example, the word 'that' in English can either be a demonstrative determiner ("Put that hat over in the box.") or a marker for a relative clause ("The hat that my cat likes to sleep on is covered in fur.") This lexical ambiguity does not have a similar ambiguity in French – the first 'that' would be translated as *ce*, while the second would be translated as *que*. Such ambiguity makes it difficult to develop a bilingual dictionary to translate English words into their corresponding French words.

For many languages, a simple one-to-one dictionary will cover large fractions of the vocabulary. For those words with multiple translations, many of the ambiguities can be resolved by looking at the context in which they appear. For the example of 'that' above, the demonstrative determiner appears as part of a noun phrase (NP) while the relativizer separates a noun phrase from a relative clause. It is relatively easy to generalize the notion of a single correspondence to multiple correspondences by developing multiple one-to-one correspondence sets and selecting among them on the basis of context.

The METLA systems use a multiple dictionary formalism to produce a context-dependent dictionary for lexical selection. Every grammatical context (defined, of course, in terms of the grammar developed in the preceding section, which is unfortunately different between the METLA-1 and -2 systems) carries with it information about which of a fixed number of dictionaries is to be used. Within an NP, then, the translation system will use a dictionary in which the lexical entry for 'that' is *ce*, while using a different dictionary with a different entry when translating relative clauses. As another example, the English word 'one' translates to the French word *une* in the context of a feminine noun phrase, *on* when alone in the subject of a sentence, and *un* in other contexts. By developing three separate dictionaries, this word can be correctly translated in all three contexts.

Further sophistication has been added by the incorporation of ϵ (epsilon, or the null string) as an additional lexical type in all languages. This allows words to be deleted (translated to ϵ) deleted in some contexts, or for ϵ to be translated to another word in specific contexts to insert words as appropriate.

20.4 Tuning by parameter optimization

Implicit in the above formalisms is the notion of describing them by parameter sets. For example, each of the several dictionaries can be seen as a function mapping words (or ϵ) to other words or as a function mapping numerical tags to other numerical tags. Each domain element can be individually mapped and changed to fit the bilingual data. Similarly, the choice of dictionaries can be described in numerical terms—in such and such a grammatical context, use dictionary number three. The end result of such description is a large number of relatively independent and tunable parameters which collectively describe a transfer function between the source and target language.

Setting the parameters at random, of course, will typically result in complete gibberish. However, by translating the source sentences in the database and comparing the translated results with the "correct" translation also listed in the database, one can produce a measure of the relative fitness of a given parameter set. Standard optimization techniques can then be applied to maximize the fitness of the parameter set.

Both systems use a standard optimization algorithm called simulated annealing[39, 48]. This technique was originally designed as a model of crystal growth and metal annealing, but has found widespread use in a variety of contexts and has the advantage of being well-known, well-studied, and reliable. In simplest terms, it is a variant of a random walk through the event-space of interest. At each step, the algorithm considers a random change to the set of parameters and measures the quality of the translations produced by the changed set. If the changed parameters result in improved performance, the system accepts the new parameter set for further work. Even if the parameters reduce performance a bit, the system *may* still accept the new parameters as long as the performance loss isn't too great. Formally speaking, the probability of acceptance is related to the Boltzmann distribution of particle energy, with performance loss equated to higher energy. As the algorithm progresses, the notion of "too great" is gradually tightened (the temperature of the system is reduced) until the algorithm accepts only improving moves and eventually will find the global performance maximum.

As all of the parameters are described in symbolic or numerical terms, the changes are simple to define and implement. The available operations for the current METLA systems are :

- The translation of a (random) lexical item in a (random) dictionary may be changed to a new item.
- The dictionary to be used in a (random) context may be changed.
- The permutation-restructuring of a particular syntactic context may be changed by swapping any two items. For example, a transfer may be changed from leaving the preposition and noun phrase in that order to reversing them. This change is equivalent to stating that the target language has postpositions instead of prepositions.
- (METLA-1 only) The set of marker words separating two non-terminals in a grammar rule may be changed by the insertion or deletion of a single lexical item.
- (METLA-1 only) Any single non-terminal in the expansion of a grammar rule may be replaced by any other non-terminal.
- (METLA-2 only) The syntactic category expected to the left of a (random) lexical item may be changed or eliminated.

- (METLA-2 only) The syntactic category expected to the right of a (random) lexical item may be changed or eliminated.
- (METLA-2 only) Similarly to the two operations above, any syntactic category expected in any subcategorization frame may be changed or eliminated.
- (METLA-2 only) The lexical category resulting from the conjunction of several phrases may be changed to another category.

The operations which are not currently permitted, although a system could in theory incorporate them include such operations as changing the number of productions in a grammar, changing the maximum number of lexical items, changing the number of dictionaries, and so forth.

21 Engineering considerations

Although simulated annealing produces good results, there is no particular psycholinguistic reason for its use. Numerical optimization is a long-studied problem and there are many other algorithms which have been developed. Early prototypes of the METLA-1 system used a variant of genetic algorithms[6], where the system maintains a large pseudopopulation of parameter sets and crosses and mutates them in a simulation of evolutionary selection. Another optimization technique, called tabu search[24, 25] (primarily a variant of hill climbing with momentum) has also been studied. In all cases, the results have been comparable or slightly worse than the results obtained from simulated annealing, and because of the size of the parameter sets and the large dimensionality of the search space, simulated annealing wins on the purely engineering consideration of efficiency.

Hidden in the description of the optimization task above is the idea of measuring the quality of the translations produced by a given parameter set. For obvious reasons, this measurement must be done numerically and automatically by the computer. Automatic measurement of translation performance is unfortunately difficult to perform. Many psycholinguistically plausible measurements are computationally expensive or technologically impossible. However, the sort of tasks that are computationally viable may produce "false positives" which appear to be related to the correct translation but in fact are very different. (Consider the effect of adding or deleting a 'not' from an English sentence.) After several experiments, the system uses a modified greatest-common-subsequence formalism[52], which should be familiar to most UNIX programmers as the *diff(1)* algorithm. Specifically, this measures the number of changes (insertions or deletions) that distinguish the translated source sentence from the desired target sentence. Sentences are thus graded on the number of words that would need to be added to or deleted from them to produce the exact form in the examples, an approximate measurement of the amount of work a human editor would need to do. Again, this measurement is not psycholinguistically motivated or defensible but has been empirically selected for engineering reasons from a larger set of candidates.

22 Simplifications and assumptions

Because the current METLA system is a research prototype, there are several assumptions and simplifications that had to be made in the course of design and development. These assumptions are in addition to and much stronger than the similar assumptions presented in section 11 which apply to any corpus-based linguistic system.

One example of such a limitation of the METLA system is the assumption of sentence by sentence translation. This results in the automatic loss of any context longer than a sentence. The difficulties this can cause can be most easily seen in so-called "pro drop" languages, languages where pronouns that can be easily inferred from the context can be dropped. In some cases, the missing pronoun can be inferred from verb morphology or a similar agreement construct. In other cases, though, there is simply a loss of information. An example of this can be demonstrated easily in Japanese : The question "*Wakarimashita ka?*" means, literally, "Understood?" Depending upon context, this can mean "Did you understand?", "Did he understand?", "Have they understood?", or several other possibilities. The correct English translation thus depends on factors external to the Japanese sentence itself. In other cases, sentences and fragments which do not parse correctly may be incorrectly translated. More general systems, such as [62, 63], which use a "fragment database" of varying lengths and syntactic complexity, may be better able to deal with these problems.

Another assumption of the METLA system is that words can be translated as individual units. This assumption may introduce two sorts of errors and limitations. First, as there is no morphological processing, the system will not necessarily identify the similarities between 'talk,' 'talked,' 'talking,' and 'talkative.' Similarly, as words are translated individually, multiword phrases may not be handled correctly. To a small extent, these may be handled by the development of individual syntactic contexts (the French translation of 'new' is *nouveau* except when capitalized and followed immediately by the word 'York'), but this would quickly result in an explosion of potential syntactic constructs beyond the ability of a reasonably sized system to handle.

Finally, the formalisms developed in the preceding sections assume certain maximum sizes, such as a maximum number of lexemes for both the source and target languages, a maximum number of production rules, a maximum fanout per production rule, and so on. Some of these assumptions are easily justifiable—for instance, a million-word corpus will have a maximum of one million different lexical tokens and in practical terms will have many fewer. By assuming a maximum number of grammatical rules, however, the expressive power of the source grammar is greatly reduced.

23 Description of experiments

23.1 Experimental framework

The standard procedure for most modern learning systems (e.g.[45, 18, 40] among many others) is to produce two separate sets of data, a training set and a testing set. The system is trained to some criterion, usually measured either in terms of performance or else a set number of training epochs, and then the actual performance measurements are taken on novel data to which the system has not been exposed. This prevents the system from merely memorizing the input data and provides a better measure of learning performance, but also requires that the researchers acquire two sets of data. In the case of METLA, this would of course be two similar aligned corpora on the same language pair, or more simply two sections of the same corpus.

Because both the METLA systems, as defined in parts VII and VIII, use simulated annealing as their primary inference mechanism, there is no practical possibility of "incremental" learning to a desired criterion, and the systems are trained over a fixed number of epochs as designated in the annealing schedule. The actual learning is run multiple times to prevent conspiracies in the random numbers from having an undue effect on the final outcome.

23.2 Corpora

The corpora used in the experiments come from a variety of sources, reflecting a variety of language types, corpora types, and syntactic complexity levels. The following are the corpora used :

Urdu An English→Urdu text taken from [77]. This book is one of many such books written in the early 20th century for the benefit of British officers in various colonies and provides a quick and easy introduction to Urdu for native English speakers[9]. The training corpus consisted of a vocabulary list and the set of example sentences (and their translations) taken from lesson 2, while the testing corpus was the set of exercises (which were translated by hand and confirmed by a native speaker of Urdu). Typologically, Urdu is an Indo-European language with a heavy influence from Arabic (as well as from other Indian languages). Structurally, it has basic word order SOV, postpositions instead of prepositions, and no definite/indefinite article distinction.

The training corpus itself consisted of thirty-five paired phrases, containing sixty-two and fifty-one English and Urdu words, respectively. The total vocabulary consisted of twenty-nine words in each of the two languages. The original test set consisted of thirteen complete sentences taken from the exercises to that lesson, but as will be discussed later, was extended to general input at the hands of a native Urdu speaker, consisting of several hundred sentences.

French An artificial English→French corpus designed to test the performance of the system on a small vocabulary but with greater syntactic complexity than the Urdu corpus. The training set consisted of forty-three sentences (containing a total of nearly 200 words) selected from a thirty word French vocabulary (corresponding to a twenty-seven word English vocabulary) and included gender distinctions, embedded relative clauses, words with ambiguous translations, reflexive and non-reflexive verbs, and multiple subcategorizations of verbs. The testing data consisted of twenty-nine similar sentences produced by a different experimenter using the same vocabulary. All translations were confirmed by a native speaker. Typologically, French is an Indo-European language with the same basic word order and structure as English, but a more pronounced gender agreement system.

Spanish A natural English→Spanish corpus extracted from children's literature as translated by professional translators[59, 10]. The testing set, in turn, came from a similar child's book by the same author[60]. The training set consisted of all sentences under a threshhold length of seven words; the testing set consisted of all sentences comprised of words found in the training set and of the same threshhold length. Typologically, Spanish is an Indo-European language with the same basic word order as English, but Spanish makes much freer use of dropped pronouns when they can be inferred from context.

The training set itself was forty-two paired sentences, containing, respectively, 183 English and 191 Spanish words. The vocabulary for this experiment comprised ninety-two English

[9] Aravind Joshi, personal communication of 15 September, 1994

and 113 Spanish words. It should be clear from this number that most words appeared only a few times (perhaps only once) and thus that the data obtained may be rather sparse, a characteristic of real data. Because most of the words are story-specific content words, only a few of the sentences in the testing story could be formed from the words in the training data, and thus only six sentences could be extracted to form the testing set.

Part VII
METLA-1

24 Introduction

METLA-1 was an experiment designed to test a direct instantiation of the METLA schema to a vanilla phrase-structure grammar, specifically a context-free grammar. Using the marker-normal form formalism developed in section 16, the system infers a basic grammar, transfer functions, and dictionary from small corpora. This prototype has been tested, as will be discussed below, on the French and Urdu corpora.

25 METLA-1 parsing algorithm

The parsing algorithm used by METLA-1 is a direct expression of the marker-normal form mathematics developed in section 16. Specifically, every non-terminal symbol (of which there are a fixed number, defined at compile time) is associated with a production rule in a modified marker-normal form. The primary modification used is that sets of marker words, instead of individuals, are used to separate the various constituents. A secondary modification is that marker words are attached, for purposes of further parsing, to the constituents that follow them. For example, a sample rule for English might be

Sentence → NP aux V det NP

where 'det' is any of the set of {a, an, the} and 'aux' is any of the set of auxiliary verbs {be, have, will, can,...} in any of their inflected forms.

Formally, the grammar can be characterized as a fixed set of rules, numbered from zero to $N-1$. Each of these rules has a fixed fanout of non-terminal symbols k, so every rule in the grammar is of the form

$$A_i \rightarrow A_x m_{i,1} A_y \cdots m_{i,k-1} A_z$$

In this notation, each $A_?$ is a non-terminal in the set $A_0 \ldots A_{N-1}$ and each $m_{?,?}$ is a set of marker words that marks the separation between the various constituents of A_i.

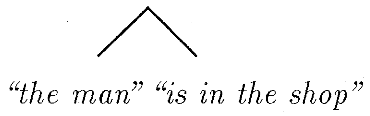
Parsing is done in a rather simplistic fashion. A_0 is by fiat designated as the starting symbol of the grammar, and the training sentences are parsed in a strict top-down fashion. Each sentence is partitioned into its constituents at the appearance of the leftmost element of each marker set, in order of appearance in the rule of grammar. For the sample rule above, this would divide a sentence at the first auxiliary, and then at the first determiner following. These constituents are then recursively parsed in accordance with the single rule corresponding to their nonterminal, and so on, until the sentence has been broken down into only lexicalized items. These items are, of course, translated using the various context-sensitive dictionaries the system has developed. Figures 2 and 3 show an example of this parsing and translation scheme in action, using an actual grammar developed by the system in the course of the English→Urdu experiments. Each stage is described twice, once in a graphical tree format and again in a more compact text-based format where constituent boundaries are marked by parentheses.

Example sentence : *"the man is in the shop"*

Initial rule categorizes sentences as imperative or declarative and doesn't actually parse the example sentence.

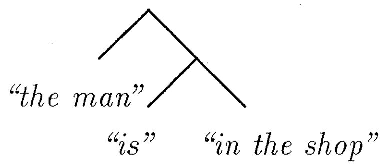
"the man is in the shop"　　the man is in the shop

Second rule divides immediately before copula of declarative sentence.



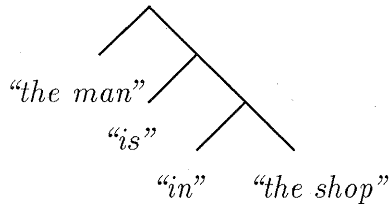
(the man) (is in the shop)

Third rule divides predicate (post-copula section) before preposition.



(the man) ((is) (in the shop))

Fourth rule divides prepositional phrase before determiner.



(the man) ((is) ((in) (the shop)))

Fifth rule (not shown) separates determiners from nouns in NP.

Figure 2: Example English parsing (derived from E→U)

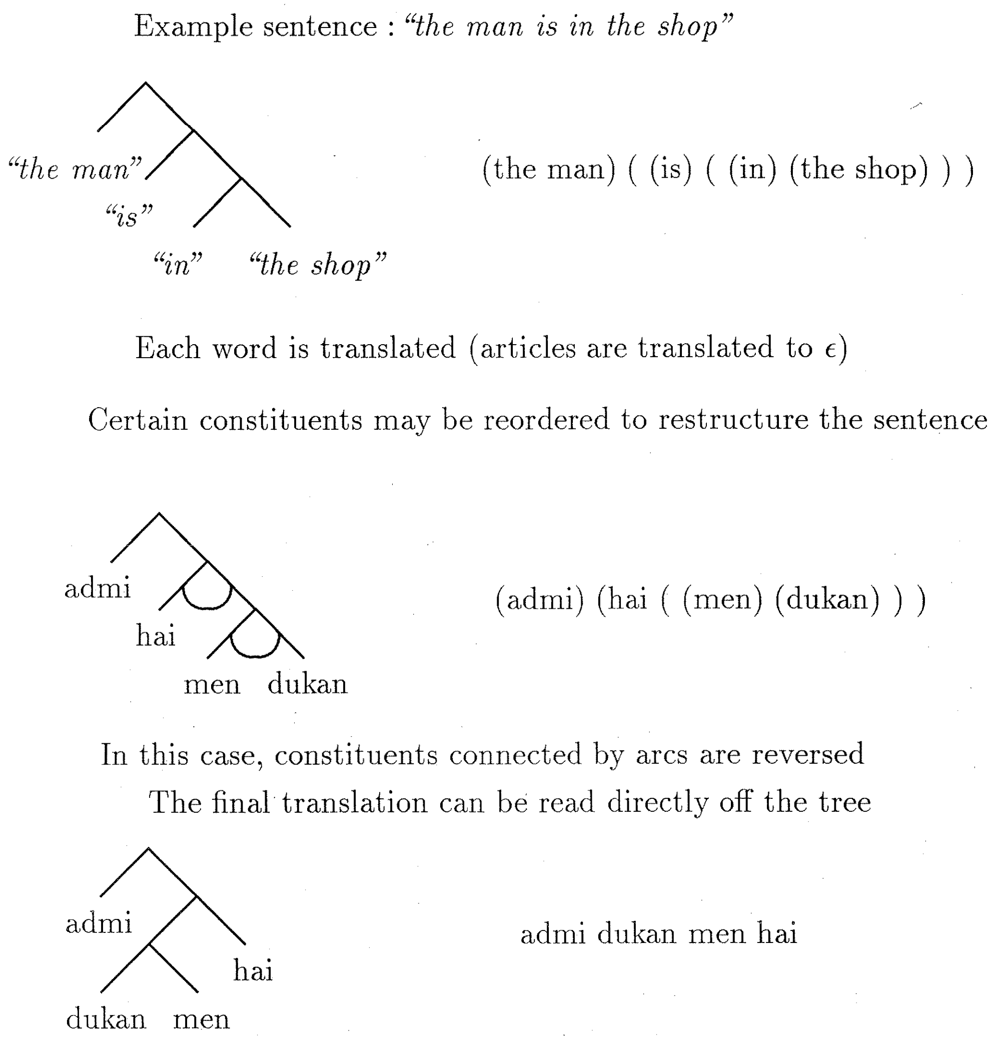

Example sentence : *"the man is in the shop"*
"the man"
"is"
"in"
"the shop"
(the man) ((is) ((in) (the shop)))
Each word is translated (articles are translated to ϵ)
Certain constituents may be reordered to restructure the sentence
admi
hai
men
dukan
(admi) (hai ((men) (dukan)))
In this case, constituents connected by arcs are reversed
The final translation can be read directly off the tree
admi
hai
dukan
men
admi dukan men hai

Figure 3: Example English→Urdu translation

The grammar comprises a set of CFG productions ($A \rightarrow BmCnD$) in marker-normal form. Each production is annotated with a restructuring permutation of the non-terminals and the number of a dictionary used to translate individual words.
The possible operations on each rule are :

- The translation of a (random) lexical item in the dictionary may be changed to a new item.
- The dictionary to be used may be changed.
- The permutation-restructuring of a particular syntactic context may be changed by swapping any two items.
- The set of marker words (e.g. m, n) may be changed by the insertion or deletion of a single lexical item.
- Any non-terminal in the expansion of a grammar rule may be replaced by any other non-terminal.

Figure 4: Parameterization for METLA-1

Because of the strict top-down nature of this parsing as well as the fact that each non-terminal is associated with only one production rule, the nature and accuracy of the grammars that METLA-1 can infer is strictly limited. In addition, because of the focus of METLA-1 only on marker *words*, many morphologically marked structures will not be found. Finally, the system starts from a randomly-chosen starting grammar and vocabulary, rather than using any techniques to select a good starting point, which slows the inference task considerably.

The complete parameterization for the METLA-1 system is shown above as figure 4.

26 METLA-1 experiments

METLA-1 was tested on both the Urdu corpus and the French corpus (described in detail in section 23.2). For both experiments, the system was presented with the training data and allowed to infer (through simulated annealing) long enough to achieve a complete "freezing" of the system. At each point, the system attempted to translate the entire training corpus, sentence by sentences, and the total amount of changes necessary to correct the translation was measured as the error. The annealing schedule used was a simple geometric progression, where every new temperature was 1% smaller than the previous temperature, and between 50 and 150 thousand changes were made at each temperature step. The annealing itself progressed over three orders of magnitude of temperature changes, typically resulting in between eight and sixteen hours of computation time on the HP Snake.

27 Evaluation of METLA-1

One of the difficulties involved with the development of a machine translation system is the evaluation of the end product. Is it better, for instance, to produce an ungrammatical translation that nonetheless seems to capture the meaning of what the original said, or to produce a grammatically flawless sentence that states something completely different from the original? How should the system respond to unusual, metaphoric, or ungrammatical inputs?

27.1 Black box evaluation

For many fully self-automated translation systems (e.g. [7]), the problem can be made worse by the relative opacity of the inferred translation system. There is no easy way to examine the internal workings of the algorithm to determine the nature and causes of a translation error or to identify how to repair the error. And for translation systems using Markov models[7] and similar oversimplified grammatical structures, it may not be possible to understand the cause of the error even after a lengthy and extensive analysis of the translation parameters, as the underlying model is too distant from people's intuitive understanding of how languages are put together.

Nonetheless, it is possible to do some sort of a black box analysis of the output of the system. Brown et al.[7], for instance, performed their analysis on the basis of hand-classification of sentences into five types, ranging from "Exact" (Identical to what the Hansard translator chose), through "Alternate" (Different phrasing but the same idea expressed), down to "Ungrammatical." This sort of hand-classification for final system evaluation is useful because it directly measures the appropriateness of the final product in a way that more automatic measures (such as diff) cannot. For the METLA-1 prototype, though, this particular classification was less useful than the classification actually used. Because of the limited vocabulary and grammar in the experiments, very few different grammatical ways to express the same idea were available. It was therefore more useful and appropriate to classify sentences (again by hand) into the categories "Correct," "Minor errors," and "Gibberish." The first category corresponds to "Exact," above. The third category describes sentences that were so syntactically ill-formed as to be unintelligible and would be a subset of Brown's "Ungrammatical" sentences. The second category would be classified by [7] sometimes as "Alternate" and sometimes as "Ungrammatical." These tend to be syntactically invalid but semantically understandable. They also tend to reflect (subtle) properties of the source language that are slightly changed in the target language. In fact, they closely resemble typical errors of first-year language students. Examples of these from the English→French experiments include deletion of sentence complementizers[10], deletion of reflexive particles, or gender errors.

When this sort of analysis is performed on the results of the English→Urdu experiments, the system learned the original training corpus (the example sentences from the lessons) perfectly and could reproduce it without errors. Testing on novel sentences (the exercises) revealed 72% completely correct, and only 7% translated as "gibberish." Upon further analysis (see below), the training corpus was shown to be unrepresentative of the test corpus, and in particular was missing coverage in context for several words. When the training corpus was updated to include coverage for the missing items, the system could still learn the training corpus perfectly and the percentage correct on novel items of the same forms increased to 100%.

The English→French experiment, because of the higher syntactic complexity in conjunction with the limited scale of the prototype, performed less well overall. Typical performance for the system on the training corpus was approximately 61% correct. On the test data, performance

[10] The 'that' in the English sentence "I believe (that) rocks sink" is optional. The corresponding 'que' in its French translation is required.

Table 1: Results from black-box analysis of French experiments

Category	Training	Testing	Limited
Exact	61%	36%	41%
Minor	29%	21%	19%
Gibberish	10%	44%	41%

was lower, with only 36% correct and a full 44% gibberish. However, when the test sentences that presented structures unrepresented in the grammar were excluded, the performance improved, up to 41% correct. These results are summarized in tabular form in table 1.

In general, comparisons of independent NLP projects, particularly independent machine translation projects using different language pairs, is not considered to be especially meaningful. Subtle differences, such as the amount of linguistic difference between the source and target languages, may make a tremendous difference in the final results. Even projects using the same language pairs may be rendered incomparable by differences in the corpora or the evaluation metric. For example, the Serbo-Croatian experiment by Koncar and Guthrie[40] used a training set of over 1/3 of a possible set of 24,000 sentences using a very limited grammar and vocabulary, but could achieve a very high (98%) accuracy rate in their translations. Without directly replicating their neural network, it is not clear what its performance would be on a more reasonable number of training sentences (relative to the grammar to be inferred).

The IBM Markov chain experiment[7] is more closely comparable, but still difficult. Here, the major difference again lies in the corpora. As discussed earlier, their system uses sentences a natural corpus, the Hansard parliamentary proceedings, as both their training and testing set. Both sets are much larger and more syntactically complex than the corresponding sets for METLA, meaning that significantly more information is available for solving a considerably more complex task. Similar difficulties in comparison obtain from the handwaving done in both projects in the hand-evaluation of the final translation project. Even where the rules are clearly stated, as above, judgements of a particular (translated) sentence about acceptability may vary from reader to reader. Even keeping this comparison difficulty in mind, however, the raw numbers from the IBM experiment are encouraging — an early version of the system was able to correctly translate 48% of the test data based on a much larger training (and testing) corpus.

Clearly, much additional work will be required before METLA turns into a commercial-quality translator. However, given the known structural limitations of the implementation and the small grammars that it used for these experiments, these still represent a significant accomplishment in the development of a psycholinguistically plausible MT system. Perhaps equally significantly, to convert the system from one language to another required approximately an hour of human effort to type in the training data, and no system modifications. This indicates that language-independent induction of transfer functions may be a viable approach to machine translation.

27.2 Glass box evaluation

A major advantage of a psycholinguistically plausible approach is that, if properly done, the output of the system can be directly converted into a grammar and dictionaries for the appropriate languages. This makes it possible to directly analyze the plausibility and appropriateness of the various transfer rules and to improve them by human intervention. Some of the simplifications made in the course of developing METLA-1 have made it more difficult to perform this task, but one can still examine the source grammar and transfer functions which the system developed and use this information to change the transfer rules or training data.

For example, in the English→Urdu experiment, the training data consisted of copula-locatives ("the hat is on the chair", "the man is in the shop") and imperative sentences ("wait in the office," "send the knife to the house"). Each of these had to be rearranged into verb-final form, and the prepositions had to be converted to postpositions. In addition, all the determiners ('a,' 'the,' 'this,' etc.) needed to be deleted, so the final result of translation would be something like the word-for-word translation of the string *"knife house to send."

Upon examination, the word classification and translation methods make sense. For an example, one of the early experiments initially divided all sentences into two parts based on the first appearance of a determiner or preposition. This divided imperatives ("wait in the office") into their verb components followed by one or more arguments which were translated by another set of rules. The translation of the verb was permuted to follow the rest of the sentence, giving the necessary verb-final form. On the other hand, declarative sentences ("the book is on the table") are passed through this initial rule unchanged, to be divided later at 'is' into subject, verb, and location, and permuted appropriately. This sort of analysis can be carried out to any desired level of detail.

Even this simplified analysis, however, is enough to demonstrate the advantage of a psycholinguistically plausible and symbolic representation. The statement "to be divided later at 'is'" is, in point of fact, slightly inaccurate. Using the first version of the training data, the system accurately inferred that 'is' serves to mark the boundary between subject and verb. However, it also inferred (wrongly) that 'knife' and 'man' were also part of that same marker group. This resulted in a small number of incorrect translations of the testing sentences.

Further examination of the input corpus showed the reason that these errors had been made. Although the system was presented with a full vocabulary list ('man'/'admi', 'house'/'ghar', and so forth) of individual words, only a subset of those words had been presented in the context of a phrase or sentence. Although the system, then, had learned that 'man' translated to 'admi,' it had no evidence about the part of speech of 'man.' The system had no way of knowing, for example, that the word 'man' was not an alternate form of the copula. In general, the lists of marker words are obviously of one or more grammatical classes, with potentially a few outliers that represent words that have never been seen in that context and therefore may or may not be relevant. With this observation, it becomes/became obvious that the input examples were not representative of the testing data, and that some new input was required. After adding two more sentences to provide context for these words, the percentage correct increased in later experiments to 100%.

Similar analysis can be done for the more complex English to French experiments Because of the greater syntactic complexity, the system as built proved to be oversimplified in several important regards and some errors were in that sense inevitable. For example, consider the following set of sentences :

(the man) (kisses) (the woman)
(le homme[11]) (embrasse) (la femme)

[11] The process that converts, for example, 'le homme' into 'l'homme' is almost purely phonological and was ignored for simplicity in the French corpus.

(the man) (kisses) (her)
*(le homme) (embrasse) (la)

*(the man) (her) (kisses)
(le homme) (la) (embrasse)

In the first pair of sentences, the pattern subject–verb–object is used in both French and English, so the identity permutation is appropriate. In the second and third pairs, the pattern becomes subject–object–verb in French, so the identity permutation is no longer appropriate. However, as each non-terminal symbol (sentence, in this case) has only one rule and one permutation associated with it, the system is forced to select one and only one of object-final or verb-final structure. These parsing errors, collectively, produced most of the marginal translations.

On the other hand, the system correctly learned appropriate translation structure for a large part of the input corpus. For example, the original sentences are parsed into three pieces based upon the existence first of a verb, and then of a determiner or pronoun. Noun phrases (which begin with a determiner in the input corpus) are themselves partitioned into classes of masculine/feminine noun phrases so that the gender of the determiner is correctly set.

The major error made by the English→French system was that it found a local maximum in reusing one of the production rules. Because any translation system should allow for recursive structures ("John said that Mary told him that Susan said that ..."), the system is permitted to call rules that have already been called. The system tended to find a local maximum where the rule used to separate masculine from feminine nouns was the same rule used to parse the original sentence, and so it conflated the two categories of verbs and feminine nouns. This meant, in turn, that sentences such ase "that woman washes a car" were divided not as "(that woman) (washes) (a car)" but instead as *"(that) (woman washes) (a car)." The robustness of this error is somewhat surprising, as it appeared in numerous experiments and was responsible for most of the gibberish errors produced by the system. Exact reasons for this robustness have been left to be explored in future research, as, in practical terms, it was felt to be more appropriate to improve the system overall than to overanalyze a deeply-flawed system. This error could presumably be rectified by allowing the system to use more production rules, but is more appropriately solved by a better parsing algorithm in general.

Some sample results are attached as tables 2 and 3. Each table shows a number of sample sentences (in the nearly opaque parenthesized format) along with their primary division into constituents. the translations of those constituents, and the final translation after it has been permuted and concatenated.

The errors in table 3 should be explained. First, note that the division of the third sentence is incorrect—"the man that touches the car" is an entire component and the main verb of the sentence is the *second* token of 'touches.' This is an artifact of the admittedly broken METLA-1 parsing algorithm, which divides at the first appearance of a given token. That this sentence is correctly translated at all is a tribute to the remarkable structural similarity between this sentence and its French translation. The fifth sentence is an example of a so-called "reflexive" verb; the proper translation should be "ce chat se lave," where 'se' is a general pronoun meaning 'self.' In English, certain verbs can be intransitive when the subject and object of the verb are the same—for example, "I shave (myself) every morning," "I wash[12]," and so forth. Some of these verbs, in turn, *must* be expressed with the reflexive particle in French but with an

[12] In some dialects, not including the author's, this concept would be expressed as "I wash up."

Table 2: Sample English→Urdu translations with partial analysis

bring the letter from the shop (bring) ((the letter) (from the shop)) (lao) ((chitthi) (dukan se)) chitthi dukan se lao
wait in the office (wait) (in the office) (thairo) (daftar men) daftar men thairo
put the box on the table (put) ((the box) (on the table)) (rakho) ((sanduq) (mez par)) sanduq mez par rakho

Table 3: Sample English→French translations with partial analysis

the glass touches a car (the glass) (touches) (a car) (le verre) (touche) (une voiture) le verre touche une voiture
she washes a cat (she) (washes) (a cat) (elle) (lave) (un chat) elle lave un chat
the man that touches a car touches a glass (the man that) (touches) (a car touches a glass) (le homme qui) (touche) (une voiture touche un verre) le homme qui touche une voiture touche un verre
that man washes a car that she creates (that man) (washes) (a car that she creates) (ce homme) (lave) *(une voiture qui elle creee) *ce homme lave une voiture qui elle creee
this cat washes (this cat) (washes) () (ce chat) (lave) () *ce chat lave

ordinary direct object otherwise. This leads, in turn, to another example of the multiple-necessary-permutation problem discussed above.

The fourth sentence is more interesting. The word 'qui' in the fourth example sentence is a relative pronoun used only for people (like 'who'). As an inanimate object, "a car" should have taken the relative pronoun 'que' as a translation of 'that'. However, notice should be taken of the mistake that the system did not make. The other token of 'that' in the sentence was a demonstrative determiner, which was correctly translated as 'ce', taking into account the gender of 'man'. The system correctly identified the second 'that' as a relative pronoun and not a demonstrative determiner. Similarly, the third sentence indicates an ability to distinguish between feminine nouns ("une voiture") and masculine ones ("un verre"), a relatively subtle grammatical point. These results, then, indicate an ability on the part of METLA-1 to determine remarkably small grammatical structures and to appropriately account for and to produce them as needed in the translation process.

27.3 Summary

Despite the evident limitations of the formalism, the results from the METLA-1 system were promising. The system demonstrated an ability to identify useful and psycholinguistically plausible structural regularities in bilingual corpora, and in particular identified such syntactic constructions as noun phrases, prepositional phrases, and verb phrases. It further identified the relationship among similar roles such as subject and object and correctly found a method of restructuring between two disparate languages.

Furthermore, the system could distinguish between multiple senses and uses of the same word(s), either within a language or across multiple languages (such as the gender distinctions in French). Finally, the system produced these results purely on the basis of examining the bilingual corpus and did not have to be specifically modified to handle a particular language category.

On the other hand, some problems were clearly apparent with the system. First and foremost, METLA-1 cannot handle multiple productions per nonterminal symbol, resulting in a tremendous loss of expressive power in the inferred grammars. Second, because of the greedy parsing scheme, the system failed to identify embedded clauses or handle many forms of recursion properly. And, finally, the system as designed cannot handle vocabularies larger than 31 words, so scalability is nearly nonexistent. The next version of the system, METLA-2, was designed to overcome these limitations both with more general data structures and a more powerful and psycholinguistically plausible parsing formalism.

Part VIII
METLA-2

28 Introduction

METLA-2 is an extension and improvement in several ways over the original METLA-1 system. The major difference, described in the next section, is the use of a different formalism, based on Categorial Grammar, for parsing and analysis. In addition, the system has been expanded and improved to deal with larger vocabulary and sentences. Furthermore, because of the strongly lexical nature of this representation, scaffolding has been added to support other sources of linguistic information such as correlation dictionaries and parts-of-speech lists, as will be discussed in part IX.

29 METLA-2 parsing algorithm

The new parsing algorithm is based directly on the Categorial Grammar formalism, as expressed by [69]. Every lexical item is assumed to have one or more senses, each of which carries certain expectations about the syntactic context in which it occurs. Basic words such as nouns, for example, are basal elements with no expectations. English articles, such as 'the', will expect to be followed by some class that incorporates nouns and adjectives. 'Have', in turn, has at least two different senses, one followed by a noun and one followed by a verb (as in "I have lived in Boulder for five years."). A fuller explanation of the CG formalism can be found in section 7.

One subtlety not discussed in that section is the problem of how large the relevant contexts need to be. For example, ditransitive verbs (such as 'give', as in the example sentence "Mary gave her cat a toy.") have three argument slots and therefore three constituents—the donor, the recipient, and the object given. It is therefore clearly insufficient for every word-sense to be constrained to a syntactic context with only one potential neighbor to the left and one potential neighbor to the right. This also applies to merely transitive verbs in verb-initial or verb-final languages, where the two arguments can be expected to be on the same side of the verb.

For this reason, a minor modification was made to the formalism as presented in section 7. Every word-sense is associated with five syntactic categories, representing, respectively, the categories expected second-to-the-left of the word, immediately-left, immediately-right, and second-to-the-right. The fifth category was the category of the final phrase produced by concatenation of all constituents. As in the METLA-1 system, all categories were represented by integers chosen from a fixed set defined at compile-time—and in particular, category zero was chosen to represent a category that is not actually present or expected. A basal element such as a noun, then, would have all four expectations set to zero, while a determiner such as 'the' might have a categorial assignment such as $3/(0)(0)[1][0]$, where only the immediately following category is significant to the word 'the.'

Figure 5 describes the complete parameterization of the METLA-2 system.

Unlike METLA-1, parsing was not done recursively. Instead, it was done by a dynamic programming routine[13] which starts by identifying all the basal elements. In a series of successive passes, each word is then examined to see whether all of its expectations are fulfilled, and if so,

[13] The actual parsing code was written by Chris Hall.

The grammar comprises a set of CG productions ($W = n/(a)(b)[c][d]$).
Each production is annotated with a restructuring permutation and with a lexical item w to substitute as the translation of W.
The possible operations on each production are :

- The translation w may be changed to a new item.
- The permutation-restructuring may be changed by swapping any two items.
- Any of the four categorial expectations $(a - d)$ may be changed to a new category, or to a zero representing no expectation.
- The final category produced (n) may be changed to a new category but not to a zero.

Figure 5: Parameterization for METLA-2

it and its expectations are placed into a parse tree. The function ends when no change occurs between one pass and the next.

Again, an example may help. Consider the following sentence :

The chair is in the office.

Both 'chair' and 'office' are nouns and can be immediately identified as such. The word 'the' subcategorizes for an immediately following noun and creates a larger phrase 'the chair.' 'In' expects to be followed by a noun phrase (or whatever category 'the chair' parses as). The word 'is', of course, has many senses, of which one is the context of s/(np)[pp].

As in the previous system, each word is translated into a target word and the structures are successively permuted and concatenated until the entire source structure has been translated. One advantage of this bottom-up translation paradigm is that it will handle sentence fragments more gracefully – the entire utterance will simply be parsed into a unit of some other type than 's.' Similarly, disconnected fragments can be individually translated into other disconnected fragments, and if one presumes that the order of the source text was selected for functional or pragmatic reasons, then the order of the target fragments need not be altered.

Again, as in METLA-1, there is no notion of morphological analysis or of multiword lexical items, although such items can be simulated by individuated word senses for individual words as part of the parsing.

The deletion of words in the translation process has been incorporated by the addition of ϵ as a lexical token, to which words may be translated in appropriate contexts. Unfortunately, it is not nearly as simple to allow addition of words. Because an ϵ exists almost everywhere in the sentence, it is necessary to tightly control the situations under which an epsilon is permitted. In METLA-2, this is incorporated by changing the semantics of the categorial structures slightly. Words with the zero element as all four of their categorial expectations are, of course, basal elements (typically nouns). Words with only one categorial expectation (e.g. adjectives and/or determiners) are represented as follows : (N/(0)(0)[K][0]), where N and K are non-zero categories. ϵ-productions occur on words with categorial expectations of the form (N/(0)(0)[0][K])—the more distant expectation is treated as an expectation which follows an immediately-filled ϵ between the word and its constituent. This ϵ, then, is treated as an ordinary lexical item and translated/permuted appropriately. This allows the system to insert words in the process of translation, as is required to correctly translate English 'not' into the French 'ne/pas' construction.

30 METLA-2 experiments

METLA-2 has been tested on three corpora. As with METLA-1, the first experiment involved the Urdu text from lesson 2 of [77]. Similarly, the second experiment was a replication of the METLA-1 experiment involving the French text. The third was a simple, but genuinely natural corpus taken from a child's picture book called *Curious George*[59] and its Spanish translation[10]. It should be noted that the two books were published separately and independently, not as a running bilingual text, and that the translation did not necessarily reflect all the properties of the source document. The two texts were aligned by hand, and then all sentences of seven words or fewer (in both versions) were extracted and presented to the METLA-2 system as training data. The testing data, in turn, was taken from the (English) book *Curious George Takes a Job*[60]. All sentences of seven words or fewer were taken from this book, and those which used only words in the training data were presented as test material.

31 Evaluation of METLA-2

31.1 Experimental results

As above, the Urdu experiments resulted in perfect translations of both the training and the testing data. The individual lexical transfer functions can be examined by hand and proved to translate perfectly for the small vocabulary and grammar. The lexical items found were intuitively "correct" in terms of categorial structure, translations, and permutations.

The French performance, measured numerically, was marginally lower than that of METLA-1 on the same corpus, achieving 36% correct and 12% marginal translations on the testing corpus. This weakness can be easily compensated for (as will be discussed in the next chapter) by the incorporation of further linguistic constraints such as entity seeding. Glass box analysis shows that the primary weakness is, again, in the greedy and deterministic parsing scheme. This weakness is aggravated (as discussed below) by the tendency for slight failures to result in catastrophic mistranslations, and additional research is required to identify the causes and preventatives for such behavior. Further difficulties are introduced by the nonuniform distribution of lexical ambiguities. Most words, especially the nouns, have only one translation, while words like 'that' have, in some cases, four or five different categorial senses that are relevant for translation. Although expected, and certainly in keeping with the predictions of the Marker Hypothesis, this made it difficult to (for engineering simplicity) provide all words with the same number of senses while balancing the complexity of 'that' against the need for speed in inference. A further development of the system, then, should take into account this imbalance and allow for varying number of categorial frames across lexical items.

Few corpus-based machine translation experiments have been done with non-technical, non-legal corpora. This is undoubtedly due, in part, to the scarcity of available corpora (of the 13 multilingual corpora on the ECI/MCI disk, only 2 short texts are "fiction"). Certainly, a major part of the reason for this is that the major producers of multilingual corpora are governments that need to produce legal documents or companies that need to produce instruction manuals

Table 4: Sample English→Spanish translations

george was very curious jorge estaba muy curioso
he was too curious también estaba curioso
they opened the door * la
george was fascinated jorge estaba fascinado
this is george éste es jorge
where was george * casa

in several languages. However, another part is due to the perceived greater freedom for a translation of fiction rather than legal/technical documents. Certainly, the literalness of the translations of *Curious George* was considerably less than that of the other two corpora. In some cases, the difficulty was purely stylistic and can be easily resolved in context—for example, the sentence "The man was happy too" was translated into *Y el hombre también.*[14] In other cases, however, the translation represented a serious change in the semantics of the target sentences.

For example, the English text used the sentence "The man took off the bag." In context (George, a monkey, has just been captured and placed in a bag), it is clear that the desired interpretation is that George is being "undressed" by having the bag taken off, as one would take the diaper off a baby. The corresponding Spanish sentences is *El hombre lo sacó de la bolsa*, literally "the man took him/it out of the bag." In addition to changing the pragmatic focus of the sentence from the bag to George and thus depersonalizing the monkey, this sentence also describes a slightly different act; in the English version, the bag is moving, which in the Spanish version, George is moving while the bag stays in place. In legal or technical documents, such a minor change could be disastrous, while in the context of fiction, it is barely noticeable.

Despite the poor quality, for METLA's purposes, of the test data, the results from the Spanish data are promising. Only six (novel) sentences made it through the stringent selection criterion — although these six sentences, in turn, represented a corpus about 15% of the size of the training data. Of those six, three were translated accurately (50%), while one more (17%) was translated to a grammatical sentence with a different meaning than the original source sentence. Even here, the error is interesting, significant, and plausible – the testing corpus included the word 'too' only as a synonym for 'also', as in the sentence presented above. The sentence "he was too curious" was translated into *también estaba curioso*, a correct Spanish sentence corresponding to "He too was curious" or "He was curious as well." Some results from the Spanish experiments are attached as table 4.

From a psycholinguistic standpoint, the results of the Spanish experiments are disappointing. The basic word order of Spanish and English are sufficiently similar that for most sentences, and especially most simple sentences, a mere word-by-word translation suffices. For this reason, many of the sentences, especially the ones that were translated correctly, were simply parsed

[14] literally, "and the man [was], too".

as a linear sequence of atoms. For this reason, there is little incentive for a plausible grammar to develop, or for that matter for any particular construct to be classified as any particular type. This explains, in part, the surprising catastrophic translation failures exhibited in the Spanish (and to a lesser extent the French) results. Unlike METLA-1, where a sentence would be analyzed from the top down, a parsing or translation failure in METLA-2 results in an unparsed phrase that blocks the parsing of larger scale structures, so a low-level parse failures typically result in a catastrophically failed sentence parse, and accordingly great mistranslation.

This discussion indicates the somewhat surprising result that psycholinguistic constraints may work *better* on language pairs that are typologically different because of the greater variability, making it easier to tease the marked structures away from the markers. Again, this is a subject to be explored in depth in future research.

31.2 METLA-2 limitations

Again, in the course of engineering the system, certain limitations and assumptions had to be accepted. First among these is the assumption of a fixed four-place template for the lexical frames. In verb-final languages, such as Urdu, it is clear that ditransitive verbs (such as 'give' and 'donate'), which subcategorize for three items, all of which appear before the verb, will be unable to cleanly fit their expectations into the two places allowed. It is not unreasonable to expect to find a language which requires even more slots—for example, consider a hypothetical language with ditransitive verbs and mandatory marking of tense and aspect in a separate word, or a language with a special word meaning "exchange objects" that subcategorizes for two subjects and two items to be exchanged.[15] In theory, this problem can be avoided temporarily by changing the system slightly from a four-place to a six- or eight- place lexical frame, but this wastes a tremendous amount of time inferring the appropriately increased numbers of zeros, and furthermore, there is no principled way to assume that any fixed size is enough to handle all possible expectations. A better approach would be to modify the system to allow for variable-sized lexical frames and for the system to develop the minimal frame for each lexical item and sense, but this sophistication has not been incorporated.

Another limitation arises from the greedy dynamic programming algorithm used for parsing. Each word is considered to be of the first sense for which the categorial expectations are met. In particular, this means that nouns, for which the categorial expectations can be met vacuously on the first pass through the sentence, are identified almost instantly. As a result, words with multiple senses that include a noun sense (such as 'can' or 'watch'), will probably be identified as a noun regardless of the context in which they appear, and then this identification will not be reconsidered in the course of continued processing. A better approach would be either to allow reconsideration and reparsing in the event of an (identified) failure, or else to place still more emphasis on the marker words as described in the MH, and to allow words such as 'the' or 'could' to force interpretations of the words that surround them. The fundamental problem is simply that parsing is non-deterministic and ambiguous and cannot necessarily be solved by local constraints alone (hence the existence of garden-path sentences). It thus requires either tremendous sophistication or an exponential algorithm to solve properly.

An observed weakness of the CG formalism as expressed in METLA-2 is that failures tend to be catastrophic. A simple glance at the errors in table 4 shows that, while most sentences were processed correctly, the sentences that were processed incorrectly were not even close in their

[15] e.g. "[John] traded [Mary] [his watch] [for a book]."

lexical selections and produced nothing even approximating a full or grammatical sentence. Because CG is a "bottom-up" formalism, errors at the lowest level (such as an individual word that is mis-categorized) cause the entire structure to collapse like a house of cards, and further processing comes to a halt. The top-down parsing, such as in METLA-1, will at least handle large-scale structures properly, even if there are difficulties with the more fine-grained details, and thus is more likely to attempt to translate a full sentence.

31.3 Summary

Categorial Grammar, as demonstrated above, has not proven to be an instant solution to the difficulties and problems of METLA-1. This should be unsurprising, as grammar acquisition is known to be a hard problem, and CG is provably equivalent to standard context-free grammars. On the other hand, METLA-2 demonstrates that the major weaknesses of METLA-1, and in particular the problems with multiple permutations and embedded relative clauses, are not insurmountable.

Furthermore, as has been discussed elsewhere, one of the major advantages of the METLA systems is the symbolic, and therefore easily understandable, nature of the transfer functions they infer. METLA-2, in particular, produces very simple, regular descriptions of grammatical structures relevant to the translation process. As will be discussed in the next chapter, these descriptions are very easy to update, modify, or restrict to reflect other linguistic knowledge obtained from other sources.

Part IX
Extensibility issues

32 Introduction

The major weakness with many artificial intelligence formalisms and prototypes is their lack of scalability. A complex algorithm that performs well in a universe of ten items may fail miserably in a universe of ten thousand—or more likely, will still (in theory) succeed, but require more time than the expected lifetime of the universe permits. Operations research, as a field, is full of these sorts of problems. An easy example of one such is the knapsack problem. Given a set of numbers and a total, the computer is asked if there is a subset that sum to the given total. This problem is known to require time exponential to the cardinality of the set of numbers. Scalability in the face of such exponential growth is usually produced either by incorporation of techniques (such as alpha-beta pruning or beam search) that reduce but do not eliminate the growth, preprocessing for an advantageous starting point, or simply applying heuristics to the process to cut the search space. For example, if all the numbers in the set are even and the required total is odd, then the problem may be quickly solved by simple parity arguments.

One of the main weaknesses of the METLA formalism is that it suffers from exactly this sort of exponential growth. In theory, every word can be syntactically unrelated to every other and their translations may be equally independent. There are therefore an exponential number of ways that the parsing and translation parameters may be set, all of which must be searched by a naive algorithm. If one accepts that CG is a sufficiently powerful formalism to represent a useful subset of natural language and assumes that the broken parser used by METLA-2 is replaced by a slower but more functional parser (such as a chart parser), then the problem of exponential growth can be seen to be the primary obstacle to the extension of the system to corpora with a vocabulary of thousands of words. However, just as the Marker Hypothesis predicts that certain regularities will be found, other sources of linguistic information can be applied to reduce this exponential growth to what is hoped will be a manageable level.

By design, the METLA system family should be amenable to the incorporation of such techniques and heuristics, since the system is designed to use a natural, understandable representation of linguistic phenomenon that is compatible with normal human understanding of languages. With very little system modification, a user should be able to encode prior knowledge in such a way as to produce a better starting point or to restructure the inference more efficiently. The remainder of this chapter describes several sample changes to use such prior knowledge and how they can be incorporated.

33 Correlation dictionary

As hinted in part VIII, much of the many-to-many translation problem arises from the translation of marker words. Content words, particularly technical jargon, proper nouns, and loan words, usually have only one translation, especially when such things as morphological case-marking and the like are filtered out. If this one translation can be inferred ahead of time, off-line, then the inference-by-optimization task can be shortened.

For such words with a one-to-one mapping (or even one to many), this problem can be solved by a statistical inspection of the corpus. Somers and Jones[72, 35] describe a system to perform such inspection. By using a statistical measurement called Dice's coefficient, essentially a measure of correlation, they can identify the word(s) most strongly associated with a given word or phrase in the source text, as well as a numerical measure of the strength of the association. This calculation is relatively fast to perform and requires no linguistic or meta-linguistic information about the source and target texts.

This sort of information can easily be incorporated into METLA. Words with a sufficiently strong association can be placed in a separate list of "words with their translations" (a correlation dictionary) that is not subject to inference and therefore requires no time to learn. The correlation dictionary becomes merely one more component in the larger "context-sensitive dictionary" that is not, in point of fact, context-sensitive and serves to translate a select set of words whose translations do not otherwise have to be learned. "Sufficiently strong," of course, can be defined in an application-specific manner — the lower the confidence threshhold, the more words will be placed in the dictionary, but also the more likely it is that words will be placed in the correlation dictionary in error, resulting in forced errors in the final system.

A further efficiency improvement can be made by noting that most words with single translations are content words, as opposed to marker words. Furthermore, as blatantly asserted in section 7, content words cluster into a manageably small number of syntactic (categorial) frames. The entire correlation dictionary, as developed above, can thus be collapsed into a small number of "content word" lexical items and treated, for purposes of learning and parsing, as equivalence classes (e.g. all proper nouns are interchangeable until the final translation stage). This simplifies the learning task immensely by reducing the learning of content words to the learning of a small number of categorial frames, instead of learning identical sets for a huge fraction of the corpus. In other words, the entire correlation dictionary can be condensed into a few lexical place-holders which describe the categorial expectations and permutation relations of an entire class of content words (such as adjectives, transitive verbs, feminine nouns), which will then be explicitly disambiguated at the time of translation. If the normal ambiguity-resolution process in the parsing routine is sufficient to separate these classes, then a single lexical placeholder will suffice, and the learning process can be sped up immensely by learning most of the vocabulary (up to 99% of it, according to [68]) as a single item rather than individually.

Finally, this correlation dictionary is not necessarily static, and can be updated as necessary to reflect new lexical items or to delete old words that were placed in the dictionary erroneously. It should also be noticed that such an improvement, if properly done, cannot result in a net loss in system performance. If a word is translated incorrectly, it can merely be removed from the dictionary and learned normally. In the extreme case of a language pair with no unidirectional one-to-one mappings between lexical items (an extremely unusual situation), then the system performs exactly as the original METLA system.

Correlation dictionaries, as alluded to in the previous chapter, have already been incorporated into the METLA-2 system. The system as designed will look for a correlation dictionary and use it if it is found. If not, the system performance is exactly the same as in the previous chapter. If one is available, all words in the dictionary are treated (for now) as one superordinate lexical class and annealed. as an group.

Because simulated annealing is not incremental, there is no direct way to do a head-to-head comparison of learning speed with and without the correlation dictionary. Instead, the system can be annealed for a much shorter period (for these experiments, approximately 3% of the full learning time) and the performance at this intermediate schedule measured against the performance of the full schedule. This has been done for the English to French experiment, with marginal results. The results from the incorporation of a small correlation dictionary (about 12% of the lexicon) are approximately 2% better than the results from METLA-2 alone with the same annealing schedule. Even so, this shows that if a correlation dictionary is available or can be cheaply obtained, it can be added to the basic METLA-2 system with little effort and some improvement can be gained.

34 Novel word support

In one sense, the acquisition of novel words is inherently unscalable. Every novel lexical item in the source language is potentially a new lexical item (which must be explicitly acquired) in the target language as well. Morphological analysis may be able to provide some guidance—for example, the (novel) inflected form of a known word may be related to a known inflectional paradigm, and so a system can guess at the correct translation. There is, however, no provably useful shortcut for the acquisition of lexical items.

It should, however, be relatively easy to produce accurate guesses about the grammar of new lexical items. Especially in METLA-2, where the grammar is stored completely as categorial frames, it is easy to sort and classify words and word-senses into various categories and thus to recognize that words such as "red," "tall," and "new" share some grammatical features. Because the set of such categories is finite by construction, it is at least theoretically possible to check every possible classification for a novel word and determine which one produces a parsable structure rather than gibberish.

Furthermore, there are many minor efficiency improvements that can be made to this naive algorithm. As Klavans pointed out[16], most novel words are nouns, so simply guessing that each novel word is a noun (with no attached subcategorization) is a surprisingly robust and accurate strategy.[17] For more sophistication, categories can be sorted by relative frequency and the more likely candidate senses can be tried first. Finally, if there are multiple senses, all of which produce a valid parse (such as is likely in a language with gendered nouns), the system can query the user "which word is this one more like?" and present a prototypical example of each of the possible senses.

A similar technique could be used to find the minimal change necessary to produce a parsable sentence, as one would expect if the system were presented with a known word in an unknown context/sense. Finally, it should be noted that knowledge of this sort, once obtained, is structurally identical to the knowledge inferred in normal operation and can be added without modification to the knowledge base.

35 Entity seeding

The preceding two sections have discussed techniques that can be used to improve the scalability of the translation process itself, either by simplifying the development of the dictionaries or by learning novel words (and their translations) within an existing grammatical framework. Another important improvement that can be made is to improve the grammatical inference by providing direction or a better starting point.

[16] in lecture, at the 1995 Annual Meeting of the *Linguistic Society of America*

[17] Dr. Menn has pointed out that this is analogous to weather prediction in Los Angeles. When in doubt, guess "sunshine."

Solomon and Wood[69], in their work on the inference of CG descriptions of language, found it necessary to seed their system with examples of primitive categories, and in particular, with nouns and noun phrases (N and NP, representing proper nouns). In this work, these are grouped together under the vague heading of "entities," constructs that denote simple, physical objects in the world. Solomon and Wood cite as justification and support for this step the psycholinguistic evidence from child language acquisition that "children learning a language do so by learning the names of things first, then how those names can be fitted together into a sequence of propositions."[69, pg. 125] A similar step can easily be taken for the METLA-2 system. By providing the system with a list of words (not necessarily complete) which are known a priori to be nouns, the system should in theory be helped in two important ways. First, because of the reduced search space, the system will be able to infer the remainder of the translation functions faster and more accurately. Second, the psycholinguistic plausibility of the grammar and translations should be increased by eliminating some potential false-positives. From a practical standpoint, nouns constitute a very large percentage of the lexical items in most languages – which means that the benefit to be obtained from such techniques is likely to be very large if a large enough list of nouns can be obtained.

Again, this conjecture has been tested with minor changes to the METLA-2 system. The English vocabulary of the English→French experiment is nearly 50% entities (between names, pronouns, and nouns). The system was presented with a list of between zero and seven words which were known in advance to be entities. The system was then allowed to learn for a short time, less than 3% of the longer runs performed for the previous chapter. Despite the immense differences in the learning time allowed, the results for the shorter learning times were similar within a few percentage points and, in fact, the results from the five word experiment were approximately 2% better than the results from the lengthy baseline experiment. These results clearly show that as simple an observation that certain words are nouns can be easily applied to the METLA systems and can result in tremendously improved learning speed.

It should finally be noted that the two improvements tested, entity seeding and the development of correlation dictionaries, are orthogonal to one another and, in fact, can complement each other. As finally implemented, the METLA-2 system can and does use both if possible to speed the learning process.

36 Grammar seeding

As in the previous section, the addition of even slight hints of grammatical information to the system can result in improved performance. An extreme case of this, of course, would be the presentation to the system of an entire (incomplete) grammar and/or translation function to the METLA system. The system could then either infer addenda to the given grammar to maximize performance, or else use the given grammar as a starting point of a (shorter) inference phase.

The first approach is problematic, primarily because of the well-known difficulties of designing a complete, accurate, and error-free ruleset to solve any major problem. Exceptions to any given rule are usually idiosyncratic and hard to pin down — or may have been ignored or forgotten in the development of the ruleset but resurface in the examination of a training corpus. The second approach is, at least at the surface, a reasonable and relatively efficient use of existing human knowledge. This is especially true given the large number of languages which have already been studied and for which written, formal, grammars exist.

In their current state of development, the **METLA** systems are not capable of making appropriate use of starting grammars. This is a simple artifact of the simulated annealing optimization algorithm. Simulated annealing begins by "melting" the system, applying random changes at a very high pseudo-temperature and paying very little attention to the success or failure of each change. From a theoretical perspective, this represents a breaking down of the (presumably) sub-optimal crystal structure present in badly-annealed metals. From a practical standpoint, however, this means that the system starts out by forgetting everything it has previously learned. If, as in the work described in parts VII and VIII, the system is assumed to begin with zero knowledge, this is not a handicap. However, if the system is supposed to add to an existing knowledge base, this is fatal. Phrased more succinctly, simulated annealing is not suitable for incremental learning.

However, this limitation arises not from any property of the **METLA** formalism itself, but from the optimization algorithm included in the two versions that have been developed. As was discussed in part VI, simulated annealing is not a necessary, or even important, part of the system but has merely been chosen as the most efficient optimization technique for this particular framework. The annealing kernel can easily be removed and replaced with another, more appropriate, optimization technique, such as genetic algorithms or tabu search. In earlier experiments, these produced results nearly identical to the results from simulated annealing, but taking slightly longer to run. Because these use the given starting point directly as a base for further improvements, they should be much more suitable for incremental learning.

More minor versions of grammar seeding, such as providing the system with a list of words known to be transitive verbs, can be handled as in the previous section—by simply encoding a known sense of those words as the final sense to be examined and not changing that sense in the learning process. General typological observations, such as a language being consistently verb-subject-object, can be similarly incorporated as a restriction on the permutation-space of some or all lexical items. Languages with a known case or gender system can be encoded directly into the grammar either as restrictions on the space to be explored or as a starting point for future consideration. In general, because of the symbolic nature of the representation, it is an easy task to encode a wide variety of useful observations, which the **METLA** system could then use to improve its processing.

37 Algorithmic scalability

Any engineer would point out that the proper way to fix an exponential algorithm, if possible, is not to make the exponential curve grow more slowly, but instead to replace it with a more efficient (e.g. polynomial) algorithm. Barring that, search space improvements should be coupled with the use of the most efficient optimization algorithm possible – which in many cases needs to be tuned to a particular problem statement and/or hardware implementation to make the maximum use of the problem geometry and the particular efficiencies of individual processors. This has not been done.

Simulated annealing is a well-known, well-regarded, general purpose optimization algorithm, but at its heart, it is simply a series of random walks through a large, ill-defined event space. Because it does not require a tremendous amount of history, it is a good technique to use on problems that require large amounts of data which much be solved in a small memory space. On a highly-parallel machine, or a machine with multiple gigabytes of physical memory, a different technique such as genetic algorithms would be a more effective use of resources and presumably result in improved performance.

Another potential improvement would lie in the tuning of the algorithm(s) to use more powerful linguistic universals. The Marker Hypothesis represents only one of many such statements about the nature of human language. Other universals, such as the well-known limitations on human short-term memory, may provide more structure to the event space, so that it can be searched more easily by simulated annealing or other algorithms.

Finally, scalability improvements are almost certainly possible by restating the problem slightly. Humans certainly do not learn language all at once, but acquire and use it over time. Even the most highly educated user may find herself face-to-face with a novel word or construction which she then incorporates into her existing language structure. Unlike humans, METLA is presented with an entire block of language from which it is expected to learn everything at once. Redefining the problem and presenting the system with small bits of language to be added incrementally would simultaneously increase the scalability of the system and provide a better tool for exploring the similarities between METLA's acquisition and the acquisition of human language.

38 Summary of scalability

Issues of scalability are very important to the eventual development of any AI system. The experiments presented in previous chapters have shown only that, given infinite time and resources, learning to translate is theoretically possible. The experiments and proposals in this chapter demonstrate that, just as computer time can be substituted for human time in the development of these translation functions, so can human time and expertise be substituted for large amounts of the necessary computation time if the expertise is available.

Because of its symbolic representation, METLA's understanding and representation of human linguistic behavior is very easy to adapt to standard linguistic knowledge, and vice versa. Almost any known problem or known constraint can be easily and productively incorporated into a system based on symbolic formalisms such as the one used in the METLA family.

Part X
Discussion

39 Improvements to the state-of-the-art

Reviewing the work presented here, it is important to take stock and realize what has and has not been demonstrated. Neither the task (example-based machine translation), the approach (inference by optimization), nor the linguistic universals applied are novel. The results are not nearly strong enough to represent a commercial threat to human translators, or even to other machine translation efforts. No more so do the results shed a strong new light on any major unsolved problems of human language cognition or language acquisition.

What is novel, first of all, is the combination of the problems of language acquisition and machine translation. Although "learning to translate," the acquisition of something that performs automatic translation, has been studied in various ways by various groups ([7, 40, 35] among others), few if any projects have involved the explicit acquisition of grammatical representations for the internal structure of the source and target corpora. This work shows that inference of such representations can be a viable part of a machine translation process and that the results can be a useful improvement upon or adjunct to more conventional example-based translation systems.

Second, this work demonstrates the feasibility of applying psycholinguistic constraints to natural language problems within a very general framework. When defined as a simple string–string mapping, machine translation requires little or no theoretical framework, with correspondingly little controversy over the type and nature of the internal representations. Seen in this light, the actual translation problem itself is irrelevant; it simply represents a "most general" natural language problem to which constraints can be added in a principled, theory-free fashion as a direct measure of their computational feasibility and flexibility.

Third, this work provides a testbed for measuring the computational accessibility of various linguistic universals, and in particular shows, yet again, the utility and power of the Marker Hypothesis. As discussed elsewhere[37], not all universals are equally useful or relevant for language acquisition, and some are downright difficult to envision how they could be acquired by children. As has been shown, the Marker Hypothesis is an easily accessible description of a universal of language that nonetheless provides strong constraints for natural language learning.

Because of the computational focus of this work, there are few direct implications to linguistics. Of course, this work represents yet another study in the validity and power of the Marker Hypothesis, this time applied to real languages. From a theoretical perspective, this work represents a potential unification traditional between phrase-structure grammars and radically lexical formalisms such as Categorial Grammar. The content/function word distinction provides a powerful mechanism both for unifying these approaches and for plugging an important learnability hole. Because so much of the syntax of natural language grammar is concentrated in the implications of a tiny class of marker words and morphemes, it may be possible to solve a large fraction of most syntactic problems by focusing on that small class.

Furthermore, the psycholinguistic implications of the reality of the Marker Hypothesis may suggest new directions for language acquisition research in general. For example, language textbooks, particularly older texts, tend to concentrate on the rote learning of lists of content words, particular nouns and verbs in their various inflectional paradigms. To the best of my knowledge, no textbook explicitly points out that such-and-such a word explicitly links or licenses a particular grammatical structure. If, however, the human language processor is explicitly set up to identify and use such marker information, then it may be a better pedagogical technique to demonstrate the marker words and their relationships first, then to add novel content words only as they become relevant to the subject(s) at hand.

In summary, then, the METLA translation systems represent, in the words of an anonymous reviewer, "a new combination of old ideas" taken from the various disciplines of optimization theory, psycholinguistics, statistics, and computational linguistics itself. Although most of the principles by themselves are old news, this work shows that psycholinguistics, specifically the Marker Hypothesis, can be useful to the language acquisition process, while computational linguistics can be useful to more linguistics by providing measures and testbeds for the properties of linguistic universals in a theory-free environment.

40 Adding new components

One of the main advantages of the METLA approach is its demonstrated ability to use other, independent, sources of symbolic linguistic information and to smoothly integrate them into the learning process as additional constraints. Chapter IX demonstrates four potential improvements that can be made by offline examination of the source and target languages. Clearly, there are many other sorts of examinations that could be made and incorporated.

Similarly, there are other categories of examination that may prove to be useful for future developments of METLA-like systems. For instance, $\bar{X}$ theory, as discussed in section 13, provides an independent set of constraints on the nature of natural language grammars. It is easy to envision modifying the skeletal grammars of METLA-1 and -2 to incorporate those constraints, and to eliminate without examination any changes that produce grammars incompatible with those constraints. In principle, any formalizable linguistic universal can be incorporated into METLA in this way. Such experiments should result first in an improved performance of the METLA system and second in a direct measurement of the practical implications and the computational accessibility of the new universals. If, for example, the results from METLA-2 + $\bar{X}$ Theory were identical, or worse, than the results from METLA-2 alone, one is faced with a strong implication that $\bar{X}$ Theory is a restatement of the implications of the Marker Hypothesis. Similarly, any improvement over the base system could be symbolically analyzed and related to the theoretical predictions and constraints of $\bar{X}$ Theory.

One of the more controversial aspects of the METLA systems has been their reliance on simulated annealing as an optimization and learning mechanic. Other than vague and metaphorical references in the literature ([47]), there have been no strong arguments propounded that children acquire language in a way that can be related to the growth of crystals upon cooling. In point of fact, few strong arguments have been propounded that make algorithmic claims at all about children's acquisition of linguistic rules. The theoretical basis for this learning algorithm, then, is not merely weak – it's nonexistent. From a practical standpoint, however, the fact that METLA works as well as it does while using a completely general and linguistically implausible learning method is a strong indication of the salience of the Marker Hypothesis, and in the absence of strong evidence in any direction, crystal growth seems as valid as any other approach. Furthermore, the system itself has been designed with an eye to addressing this very point in future work. The learning "module" can easily be extracted and replaced by a more psychologically plausible module. Now and in the future, the neural network literature in particular should prove to be a valuable source of ideas and algorithms that can be easily adapted to serve as the inference kernel of a METLA-like system.

Another interesting new component would be the development of a system capable of focusing on smaller components than mere marker words. In general, the METLA systems treat their inputs as strings of unanalyzed symbols with no relationship to one another. This is clearly not the case for real languages. Various forms of morphological relationships abound in the languages of the world. There is no theoretical reason why the METLA systems could not operate on strings of morphemes, providing that other systems were available to, on the front, convert standard linguistic text into strings of morphemes, and, on the other end, to convert a list of translated morphemes back into words and phrases. This task, again, may be the sort of task at which neural networks excel, and each individual system can be built monolingually, without special reference to its eventual use as part of a translation system.

A similar argument could be made about the incorporation of additional, separately derived, linguistic information, and in particular about the incorporation of part-of-speech tagging as a preprocessor to a METLA-like system. Part of speech tagging, of course, is one of the major success stories of the statistical natural language community; sophisticated system such as [17] can routinely tag novel text (including novel words) with their parts of speech at a success rate well over 95%. The improvements to be gained simply by identifying which words in the input corpus are nouns have already been demonstrated in the previous part. By being able to identify a large class of words as, for example, "proper nouns" as opposed to mere "nouns," the system could gain these advantages and more. Similarly, by being able to accurately lump "transitive verbs" together into a single categorial frame, the system could learn the appropriate framework and transfer functions much more easily. Even as simple a tagging set as "content word"/"marker word" could aid the system by allowing it to quickly focus the syntactic inference on the words with the most effect on the syntax of the sentence.

A final incremental improvement that should be discussed is the incorporation of probabilistic information. In a sense, this has already been touched upon in the discussion of the correlation dictionary; any word which is sufficiently strongly associated with a particular translation should simply be assumed to be given that translation. This sort of improvement, however, can be more generally applied. For instance, marker words tend to very strongly subcategorize for a given categorial framework. The number of times that the word 'the' appears in English text and is not followed by a noun is almost vanishingly small. On the other hand, verbs tend to be much more flexible in the argument structure that they permit. Such adjuncts as time, place, and manner can be considered optional arguments to the verb. For this reason, verbs could be treated as having only probabilistic expectations, with associated strengths and plausibility. This also allows a system to more accurately parse ambiguous sentences such as "Time flies like an arrow." It is unclear exactly how to incorporate such softness and flexibility, but it may be an important part of a final, working system.

Part XI
Extensions and Conclusions

41 Summary of results

The METLA systems (METLA-1 and METLA-2) are two instantiations of a linguistically plausible approach to the problem of machine translation. They both infer symbolic grammars and transfer functions from an aligned bilingual corpus of sentences and present these functions in a manner which can be easily applied to novel sentences or modified by hand to improve performance.

Both systems rely heavily on a particular linguistic universal called the Marker Hypothesis, which states in essence that "marker" words (and morphemes) play an important role in syntactic processing in all languages. The METLA-1 system uses a direct computational description of marker words within the framework of traditional phrase-structure grammars, while the METLA-2 system applies the marker hypothesis, and especially the marker/content word distinction, to Categorial Grammar. Both systems, in addition, use a context-sensitive bilingual dictionary and a set of permutation-based restructuring rules to perform the translation task itself. These components, in turn, are embedded in a standard multivariate optimization framework and allowed to learn as much as possible from a small training corpus before performance measurements are taken.

Because of the language-independence of these systems, it is easily possible to learn any language pair from any available corpus without system modifications. Experiments have been performed on English, French, Spanish, and Urdu, with degrees of success ranging from approximately 40% to 100% accuracy in translation. Although disappointing in purely numerical terms, the linguistic plausibility and understandability of the final translation systems represent a major step forward in the state of the art.

Further advantages that can be obtained from a METLA-like approach include the ability to smoothly incorporate independently derived linguistic information such as parts of speech, previously known grammars, and so forth. Two such experiments have been performed, with others outlined and described.

In summary, the METLA systems demonstrate that psycholinguistics in general, and the Marker Hypothesis in particular, can be an important guide to the general natural language processing problem. They further indicate that the content/function word distinction provides an important source of linguistic information which is all too often ignored in NLP systems.

42 Extensions

The preceding part described some of the simple, easily feasible extensions that could be made to the METLA system to result in improvements to its performance on the pairwise machine translation problem. Any or all of them could presumably be implemented and tested within a time frame of a year or less by a small group of motivated students. This part, in turn, describes major extensions to METLA that might involve a major rethinking of the entire paradigm, but which in turn may result in significant, novel breakthroughs of major proportions. Perhaps a better title for this section would be "Machine Translation and Science Fiction" or "Top Ten reasons for the author to get a job writing for Paramount."

In its current instantiation, METLA certainly represents a lower bound on human linguistic abilities, and yet can infer remarkably detailed syntactic structures. Of course, there are major holes in what METLA cannot handle, such as context, pronominal antecedents, world-knowledge, and such. These structures are more properly handled not by incremental improvement of the statistical basis for METLA, but by the incorporation of a similarly minimalist theory of mind – for example, what concepts are stressed by this sentence, and how would this other sentence be translated given that these concepts have been recently foregrounded? METLA, then, can provide a testbed for various theories of mind and of how they can be related to linguistic competence and performance.

43 Conclusions

In this thesis, I have demonstrated first that standard, data-intensive, NLP systems and in particular, statistical example-based machine translation systems can benefit from the incorporation of symbolic constraints describing linguistic universals. Furthermore, the Marker Hypothesis, as described by Green[27], is a powerful and accessible universal which can easily be incorporated into such a system.

This has been shown in the development and testing of two systems, named METLA-1 and -2, which rely on the Marker Hypothesis to guide the inference of general context-free grammars to produce transfer functions from unanalyzed aligned bilingual corpora. This system results in moderate accuracy as compared with purely statistical systems, but a greatly improved understandability and maintainability. Furthermore, the system can easily and cleanly incorporate additional sources of linguistic information that cannot easily be added to more traditional systems.

Like many projects, this opens at least as many questions as it closes. What other sources of linguistic information are appropriate for a METLA-like system? What improvements are possible? What features should be added to produce a better model of human language abilities? Do these features produce better translations? What are the specific linguistic properties that make the METLA system perform better or worse on a specific corpus pair? These and other questions represent years of incremental improvements to the basic questions framed by METLA. The METLA system is a signpost, not a destination. However, the directions it gives are potentially of tremendous importance to the future development of NLP and linguistic modelling as a whole.

References

[1] Dana Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45:117–35, 1980.

[2] J. K. Baker. Trainable grammars for speech recognition. In Jared J. Wolf and Dennis K. Klatt, editors, *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*, New York, 1979. Algorithmics.

[3] Yehoshua Bar-Hillel. A quasi-arithmatical notation for syntactic description. *Language*, 29:47–58, 1953.

[4] Robert C. Berwick. *The Acquisition of Syntactic Knowledge.* MIT Press, Cambridge, Mass., second edition, 1985.

[5] Robert C. Berwick and Sam Pilato. Learning syntax by automata induction. *Machine Learning*, 2(1):9–38, 1987.

[6] Albert Donally Bethke. *Genetic Algorithms as Function Optimizers.* PhD thesis, University of Michigan, January 1981.

[7] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June 1990.

[8] G. Carroll and E. Charniak. Learning probabilistic dependency grammars from corpora. In *Working Notes, Fall Symposium Series*, pages 25–32. AAAI, 1992. Cited in [Charniak 1993].

[9] G. Carroll and E. Charniak. Two experiments on learning probabilistic dependency grammars from labeled text. In *Workshop Notes, Statistically-Based NLP Techniques*, pages 1–13. AAAI, 1992. Cited in [Charniak 1993].

[10] José María Catalá and Eugenia Tusquets. *Jorge El Curioso.* Houghton Mifflin Company, Boston, 1990. Translation of (Rey, 1941).

[11] Eugene Charniak. *Statistical Language Learning.* MIT Press, Cambridge, MA, 1993.

[12] Noam Chomsky. On certain formal properties of grammar. *Information and Control*, 2(2):137–67, 1959.

[13] Noam Chomsky. Remarks on nominalization. In *Studies on Semantics in Generative Grammar*, pages 11–61. Mouton, The Hague, 1972.

[14] Noam Chomsky. *Lectures on Government and Binding.* Foris Publications, Dordrecht, Holland, 1981.

[15] S. Crespi-Reghizzi. An effective model for grammatical inference. In B. Gilchrist, editor, *Information Processing IFPI Congress 71*, New York, 1972. North-Holland.

[16] William Croft. *Typology and Universals.* Cambridge University Press, Cambridge, 1990.

[17] Doug Cutting, Jilian Kupiec, Jan Pedersen, and Penelope Sibun. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Lanugage Processing*, Trento, Italy, April 1992. Association for Computational Linguistics. Also available as Xerox PARC technical report SSL-92-01.

[18] Sreerupa Das, C. Lee Giles, and Guo-Zheng Sun. Learning context-free grammars: Capabilities and limitations of a recurrent neural network with an external stack memory. In *Proceedings of The Fourteenth Annual Conference of the Cognitive Science Society*, 1992.

[19] Bonnie Jean Dorr. *Machine Translation : A View from the Lexicon.* MIT Press, Cambridge, MA, 1993.

[20] Jerome A. Feldman, George Lakoff, Andreas Stolcke, and Susan Hollbach Weber. Miniature language acquisition: A touchstone for cognitive science. Technical Report TR-90-009, International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, California 94704, March 1990.

[21] Charles Fillmore. The case for case. In *Universals of Linguistic Theory.* Holt, Rinehart, and Winston, New York, 1967.

[22] Charles Fillmore, Paul Kay, and Mary O'Connor. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64:510–538, 1988.

[23] A. Gervais. *Evaluation of the TAUM-AVIATION Machine Translation Pilot System.* Translation Bureau, Secretary of State, Ottawa, Canada, 1980.

[24] Fred Glover and Manuel Laguna. Tabu search. In *Modern Heuristic Techniques for Combinatorial Problems.* Blackwell Scientific Publications, 1992.

[25] Fred Glover, Eric Taillard, and Dominique de Werra. A user's guide to tabu search. unpublished monograph, 1991.

[26] E. Mark Gold. Language identification in the limit. *Information and Control*, 10:447–74, 1967.

[27] T. R. G. Green. The necessity of syntax markers: Two experiments with artificial languages. *Journal of Verbal Learning and Verbal Behavior*, 18:481–96, 1979.

[28] Joseph H. Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Grammar.* MIT Press, Cambridge, MA, 1966.

[29] Chris Hall, Patrick Juola, and Adam Boggs. Morpheus : A tool for the lexical analysis of corpora for morpheme segmentation. In *Proceedings of the 1994 Mid-America Linguistics Conference*, Lawrence, Kansas, October 1994.

[30] John E. Hopcroft and Jeffrey D. Ullman. *Formal Languages and Their Relation to Automata.* Addison-Wesley Publishing Company, Reading, Mass., 1969.

[31] John E. Hopcroft and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley Publishing Company, Reading, Mass., 1979.

[32] W. John Hutchins and Harold L. Somers. *METEO*, chapter 12. Academic Press, Ltd., London, 1992.

[33] Ray S. Jackendoff. *$\bar{X}$ Syntax : A Study of Phrase Structure.* MIT Press, Cambridge, MA, 1977.

[34] Stig Johannsson and Knut Hofland. *Frequency Analysis of English Vocabulary and Grammar.* Clarendon Press, Oxford, 1989.

[35] Daniel Jones and Melina Alexa. Towards automatically aligning German compounds with English word groups in an example-based translation system. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, pages 66–70, Manchester, UK, September 1994.

[36] Patrick Juola. Machine translation and lojban. *Ju'i Lojypli*, 8, February 1989.

[37] Patrick Juola. Computational accessibility of linguistic universals. In *Submitted to the Seventeenth Annual Conference of the Cognitive Science Society*, Pittsburgh, Pennsylvania, July 1995.

[38] Edward Keenan and Bernard Comrie. Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, 8:63–99, 1977.

[39] S. Kirkpatrick, C. D. Gelatt, Jr., and M. Vecchi. Optimization by simulated annealing. *Science*, 20:671–80, 1983.

[40] Nenad Koncar and Gregory Guthrie. A natural language translation neural network. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, pages 71–77, Manchester, UK, September 1994.

[41] Pat Langley. Editorial : Machine learning and grammar induction. *Machine Learning*, 2(1):5–8, 1987.

[42] K. Lari and S.J. Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56, 1990.

[43] Fred Lerdahl and Ray Jackendoff. *A Generative Theory of Tonal Music.* MIT Press, Cambridge, Mass., 1983.

[44] S.M. Lucas and R.I. Damper. Syntactic neural networks. *Connection Science*, 2(3):195–221, 1990.

[45] James L. McClelland, David E. Rumelhart, and the PDP Research Group. *Parallel Distributed Processing : Explorations in the Microstructure of Cognition.* MIT Press, Cambridge, Mass., 1987.

[46] David McNeill. Developmental psycholinguistics. In F. Smith and G.A. Miller, editors, *The Genesis of Language: A Psycholinguistic Approach*, pages 15–84. MIT Press, Cambridge, Mass., 1966.

[47] Lise Menn. On the origin and growth of phonological and syntactic rules. In *Papers from the Ninth Regional Meeting of the Chicago Linguistic Society*, Foster Hall, 1130 East 59th Street, Chicago, April 1973. Chicago Linguistic Society.

[48] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–92, 1953.

[49] Risto Miikkulainen. A neural network model of script processing and memory. In *International Workshop on Fundamental Research for the Future Generation of Natural Language Processing*, Kansai Science City, Japan, July 1991. ATR Interpreting Telephony Research Laboratories.

[50] James L. Morgan, Richard P. Meier, and Elissa L. Newport. Facilitating the acquisition of syntax with cross-sentential cues to phrase structure. *Journal of Memory and Language*, 28:360–74, 1989.

[51] Kazuo Mori and Shannon D. Moeser. The role of syntax markers and semantic referents in learning an artificial language. *Journal of Verbal Learning and Verbal Behavior*, 22:701–18, 1983.

[52] Eugene W. Myers. An O(ND) difference algorithm and its variations. *Algorithmica*, 1:251–56, 1986.

[53] Makoto Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Barnerji, editors, *Artificial and Human Intelligence*, pages 173–80. North-Holland, 1984.

[54] Sergei Nirenburg, Jaime Carbonell, Masaru Tomita, and Kenneth Goodman. *Machine Translation : A Knowledge-based Approach.* Morgan Kauffmann Publishers, San Mateo, Calif., 1992.

[55] Fernando Pereira and Yves Schabes. Inside-outside reestimation from partially bracketed corpora. In *Proceedings of the Conference of 30th Annual Meeting of the Association for Computational Linguistics*, 1992.

[56] Geoffrey K. Pullum. *The Great Eskimo Vocabulary Hoax.* University of Chicago Press, 1991.

[57] Willard Van Orman Quine. Ontological relativity. In *Ontological Relativity and Other Essays.* Cambridge University Press, New York, 1969.

[58] Terry Regier. The acquisition of lexical semantics for spacial terms : A connectionist model of perceptual categorization. Technical Report TR-92-062, International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, California 94704, September 1992.

[59] H. A. Rey. *Curious George.* Houghton Mifflin Company, Boston, 1941.

[60] H. A. Rey. *Curious George Takes a Job.* Houghton Mifflin Company, Boston, 1947.

[61] George D. Romanos. *Quine and Analytic Philosophy: the Language of Language.* MIT Press, 28 Carleton Street, Cambridge, 1983.

[62] Satoshi Sato. Example-based translation approach. In *International Workshop on Fundamental Research for the Future Generation of Natural Language Processing*, Kansai Science City, Japan, July 1991. ATR Interpreting Telephony Research Laboratories.

[63] Satoshi Sato and Makoto Nagao. Toward memory-based translation. In *Proceedings of COLING-90*, volume 3, pages 247–52, 1990.

[64] Roger C. Schank and Robert P. Abelson. *Scripts, Plans, Goals, and Understanding : An Inquiry into Human Knowledge Structures.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1977.

[65] Dan Isaac Slobin. *Psycholinguistics.* Scott, Foresman, and Company, Glenview, Ill., second edition, 1979.

[66] Dan Isaac Slobin. Crosslinguistic evidence for the language-making capacity. In Dan Isaac Slobin, editor, *The Cross-Linguistic Study of Language Acquisition*, volume 2 : Theoretical Issues, chapter 15, pages 1157–1256. Lawrence Erlbaum Associates, Inc., 365 Broadway, Hillsdale, New Jersey, 1985.

[67] Jonathan Slocum, editor. *Machine Translation Systems.* Cambridge University Press, Cambridge, 1988.

[68] Tony C. Smith and Ian H. Witten. Language inference from function words. Technical Report 1993/3, University of Waikato, New Zealand, Jan 1993.

[69] Danny Solomon and Mary McGee Wood. Learning a radically lexical grammar. In *The Balancing Act : Combining Symbolic and Statistical Approaches to Language (post-ACL'94 workshop)*, pages 122–130, Las Cruces, New Mexico, July 1994.

[70] R. J. Solomonoff. A formal theory of inductive inference, part 1. *Information and Control*, 7(1):1–22, 1964.

[71] R. J. Solomonoff. A formal theory of inductive inference, part 2. *Information and Control*, 7(3):224–54, 1964.

[72] Harold Somers, Ian McLean, and Daniel Jones. Experiments in multilingual example-based generation. In *3rd International Conference on the Cognitive Science of Natural Language Processing (CSNLP-94)*, Dublin, Ireland, July 1994.

[73] Andreas Stolcke and Stephen Omohundro. Hidden Markov model induction by Bayesian model merging. In C. L. Giles, S. J. Hanson, and J. D. Cowan, editors, *Advances in Neural Information Processing Systems V.* Morgan Kaufman, 1993.

[74] E. Sumita and H. Iida. Experiments and prospects of example-based machine translation. *Proceedings of the ACM*, 1991.

[75] E. Sumita, H. Iida, and H. Kohyama. Translating with examples : A new approach to machine translation. In *The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*, 1990.

[76] Leonard Talmy. The relation of grammar to cognition. In Brygida Rudzka-Ostyn, editor, *Topics in Cognitive Linguistics*, pages 165–205. John Benjamins Publishing Co., Amsterdam/Philadelphia, 1988.

[77] Aziz ur Rahman. *Teach Yourself Urdu in Two Months.* Azizi's Oriental Book Depot, II, K, 14/4, Nazimabad, Karachi–18, Pakistan, 22nd edition, 1958.

[78] Terry Winograd. *Understanding Natural Language.* Academic Press, New York, 1972.

[79] Mary McGee Wood. A categorial syntax for coordinate constructions. Technical Report UMCS-89-2-1, Department of Computer Science, University of Manchester, 1989.

[80] Mary McGee Wood. *Categorial Grammars.* Routledge, London, 1993.

Part XII
Getting METLA-2

The METLA-2 system consists of approximately 2000 lines of ANSI C, written originally for DEC Alphas and/or HP Apollo 700's. It has since been ported to a variety of platforms, including Sun workstations, IBM-PC's, and a Kendall Square Research KSR-1/64. The source itself is fairly compact, occupying less than 100K of storage, while the size of the necessary data sets of course varies with the complexity and size of the input data, and in particular with the number of words in the translation vocabulary.

Unfortunately, space concerns as well as basic esthetic judgement prevent the inclusion of the entire METLA-2 source code as part of the printed version of this thesis. For those interested, copies can be obtained via anonymous FTP from *ftp.cs.colorado.edu* in the directory */pub/distribs/metla*. Bug reports and questions should be addressed to the author, currently at *juola@cs.colorado.edu*.

Copyright laws, unfortunately, restrict my making available the various corpora used in the development and testing of METLA. The Urdu corpus may by this time be public domain, but the copyright on the book having been renewed as recently as the mid-fifties, that's uncertain. The English→Spanish corpus, *Curious George* and *Jorge el Curioso* are, of course, protected, but easily available from any well-equipped bookstore.

The artificial English→French corpus, on the other hand, is not protected by any publishing company. Both the training and testing sets, as well as the vocabulary used to build them, are freely available in the same directory as the METLA source itself and can be distributed freely.