

**Lexical Transfer:
Between a Source Rock and a Hard Target**

Alan K. MELBY

Department of Linguistics
Brigham Young University
Provo, Utah 84602
USA

Abstract

Lexical transfer is the point of transition between an unchangeable source text (a rock) and an infinite array of target texts (a hard place to find an acceptable one). The author's Coling86 paper (pp. 104-106) described a new methodology for testing lexical transfer in machine translation. This paper reports on the application of that methodology to a test of the DLT system and describes a synchronized bilingual data base by-product. Further use of the methodology is encouraged.

Topic: Evaluation of machine translation

Additional Topic: Text data bases

1. DEFINITION OF LEXICAL TRANSFER

Although the term lexical transfer applies most directly to machine translation systems based on a linguistic model of analysis, transfer, and generation, it can also be applied to systems in which there is no direct correspondance between source and target words (such as interlingual systems) by defining lexical transfer as the point in processing where the target lexical forms first appear. It is the crucial point at which all the information available in the system must be brought to bear on the problem of choosing lexical forms. The choices must be appropriate, or all the sophistication of the system in other areas such as word order and discourse markers will be of no avail in producing acceptable output.

Lexical transfer must be based on the source text, which is generally a given. That is, one cannot come back during the evaluation of the output and suggest that the source text be changed to better match the target text. Thus the source text can be compared to a rock or to a text carved in stone. The target text, on the other hand, is supposedly somewhere in an infinite collection of texts composed of members of an infinite set of sentences generated from a large finite set of lexical items. Even if one could list in advance all the possible translations for each word, which one cannot, the task is daunting. Thus the space of all target language texts is a hard place in which to find an appropriate one. For those readers not familiar with the saying "between a rock and a hard place" on which the title of this paper is based, I mention that it refers to a difficult situation in which all apparent options present problems. Of course, the title twists the saying in several ways. Its purpose is both to emphasize the difficulty of lexical transfer and to illustrate it. The illustration is this: Please translate the title into some other language, basing it on an equivalent target language saying adjusted to describe lexical transfer. It is novel situations that make lexical transfer truly difficult to program for.

2. BACKGROUND

At COLING86, the author presented a methodology for testing the lexical transfer mechanism of a machine translation system. Since then, the proposed test has been performed on an early version of the DLT (Distributed Language Translation) machine translation system. This paper will describe the results of the test. In the course of performing the test, a bilingual data base of French and English texts was produced. This data base consists of paired documents with

synchronized paragraphs. This paper will also describe the data base, which has been edited and is now available to qualified researchers for a small fee.

3. TUNING DISTORTION

A good methodology for testing lexical transfer must avoid the trap of "tuning distortion". Tuning distortion refers to the misleading (distorted) results obtained from a machine translation system when its dictionaries and algorithms are adjusted (tuned) to a particular text. Almost any machine translation system can produce brilliant results when the same text is run through it again and again with successive tuning. The power of tuning is well-known and has been given a name in AI research, namely, defining a microworld. Corresponding to this power is the well-known difficulty of expanding a microworld system to function intelligently in a macroworld. In a machine translation system, difficulties arise when a tuned system is applied to a new text.

4. THE WORD LIST APPROACH

To avoid tuning distortion in a test of lexical transfer, one can build a dictionary from a word list without knowing what text will be supplied later, except that it will consist of words from the word list. This approach has significant advantages over supplying an arbitrary text and upgrading the dictionaries to handle the text, because there is a conscious or unconscious tuning of the dictionary entries during the upgrade process so long as the text is available.

In the word list approach, all the words of the text are combined with a number of misleading words which make it difficult to tell what is the subject field of the text. Then the combined words are sorted into alphabetical order and reduced to their basic forms. The alphabetic word list is supplied to the machine translation dictionary updaters and the dictionaries are stabilized. Then the text is provided and immediately translated without any updates to the system and without any words missing from the dictionaries.

If one argues that this method forces the dictionary updaters to consider too many possible collocations of each word in the list, one is simply complaining about the difficulty of handling real text. At least this method allows realistic testing of a system BEFORE its dictionaries have reached full size. If there is a problem in the system design, it is

better to find out with dictionaries of one thousand words and all their collocations than after the dictionaries contain thirty thousand words.

5. SOME RESULTS FROM THE DLT TEST

The DLT machine translation project is a venture of the BSO company in Utrecht. The word list approach was used to test its lexical transfer phase even before the syntactic analysis phase was complete. This was done by manually analyzing the test sentences. The four test passages included over 2000 word tokens which reduced to about 600 content word types, to which were added about 200 misleading words. The word list of about 800 words was used to build dictionaries containing thousands of entries.

After the texts were translated by the DLT system from English to French (during the first quarter of 1987), they were compared with official versions of the texts prepared by professional human translators at the CEC. This comparison revealed that many words matched the official language versions, some were acceptable synonyms and, as expected, some words were translated inappropriately. The DLT project is congratulated on the overall success of the experiment. The problem words to be discussed in the paper are not intended to be simply a criticism of DLT but rather observations that may be of interest to all machine translation researchers. Some inappropriate translations would be easily corrected by detecting predictable collocations. In the DLT test, the collocation software was not operational. For example, *computer-assisted* requires a particular translation of *assisted*. Another problem is *bring*, which can sometimes be translated as *faire venir* but which is normally translated as *prendre* in the context of the expression *bring x to y's consciousness*. This requires syntactic transfer of a type the DLT project calls metataxis and which was not implemented for the test. In a recent issue of *Language Monthly* (December 1987, p. 7), it was reported that Peter Lau, of the Eurotra project, said, at the 1987 Aslib conference, that the real problem of machine translation is not the "reduction of structural differences" but rather the "disambiguation of lexical entries". The DLT test focussed on such lexical transfer problems.

Some words of interest from the test are: *hardware, area, sheet, practice, giving, perform, produce, schedule, concern, field, application, induced, lead, benefit, cover, bachelor, course, duty, and form*.

For each of the above words, the DLT system produced a translation which was not appropriate to the context. These were not the only mistakes, but on the other hand, the DLT system translated the majority of the words (60 percent) acceptably, while a fourth (25 percent) were problems for one reason or another, with 15 percent in the gray area between acceptability and unacceptability.

The reader is invited to consider how these words would be handled in his or her system, be it machine translation, content analysis, or other natural language processing system. How would the proper distinctions be made or an appropriate translation for these words be found without being tuned to a particular text or sublanguage?

Not surprisingly, the word *hardware* needs a special translation when referring to computer hardware. But in today's technical documents, there can be reference to computers but also to hardware in a more general sense or in reference to tools or weapons. How can the appropriate selection be made without an enormous world model and a system which truly understands the text? (Shades of Bar-Hillel) Another example is the word *area*, which can be translated *région* or *partie*. However, these two options are not interchangeable and the distinction is subtle and not dependent on predictable collocations. A *sheet* can be a *drap* (on a bed), a *feuille* (of paper), or a *lame*, depending on context. Unfortunately for lexical transfer, the word *sheet* will not always be followed by a prepositional phrase indicating the composition of the sheet.

A *practice* can be what a medical doctor does when treating people, what a musician does to get ready for a concert, or what is normally done in some endeavor. These three may be translated differently.

The verb form *giving* can refer to a transfer of an object to someone or to a result ("one plus three gives four").

To *perform* can refer to one's normal duty or to a stage performance and may be translated differently. Likewise, to *produce* can translate differently, depending on whether one is talking about a play or factory.

The reader can use a standard dictionary to see the difficulties in the following words: *schedule* (time table or price list), *concern* (interest or anxiety), *field* (literal area of terrain or figurative field of interest), *application* (treatment or level of effort), *induced* (social or

electromagnetic pressure), *lead* (a wire or a sales contact), *benefit* (advantage or government payment), *cover* (lid or abstract limit), *duty* (obligation or import tax), and *course* (path or academic class).

Two of the words in the list involve an element of poetic justice. Katz and Fodor distinguished the academic degree and unmarried man readings of *bachelor* with markers, but did not tell DLT how to distinguish between them when the word is encountered in text. And the translation of *form* depends on its content.

6. THE BILINGUAL DATA BASE

Preliminary to the DLT test, a corpus of texts was gathered to assist in dictionary development. A portion of the corpus was kept secret and the test passages were chosen from this portion. A larger portion was made available to the DLT project for lexical studies. The Waterloo concordance system was used to generate KWIC listings for the lexicographers to observe the various uses of words in actual texts.

The bilingual data base used in the test was derived from public domain documents of the CEC and the United Nations, to avoid copyright problems. It consists of twenty documents, each with an English and a French version, on subjects ranging from migrant workers to the ESA spacelab to the automobile industry to agriculture. The documents were first scanned using a Kurzweil OCR device. Then the disk files were hand-edited into 400 small synchronized files, 200 English and 200 French, representing a total of about eight megabytes of data (i.e. over one million words). As of this writing, the small files are being further proofed against the original documents, and the paragraphs or other logical units of the texts are being synchronized by editing in segment number marks. These marks are used by a simple preprocessing program to produce synchronized two-column bilingual output for indexing by WordCruncher, a new dynamic concordance system which has become available since the project began. This two-column format will facilitate the study of all occurrences of words or expressions with the other language segment automatically displayed to allow the researcher to quickly see how that word or expression was translated in the corpus.

The edited corpus is available with permission of BSO for a modest fee to qualified researchers. It is hoped that the corpus, the word list methodology, and the results of the DLT test will be of use to others in the machine translation community.