# More may not be better - a discussion note on CAT dictionaries

by Robert WOOD
and
Logan WRIGHT
of
Automated Language Processing Systems ALPS

When comparing or discussing dictionaries used in computer-assisted translation systems, it should be remembered that a basic translation dictionary is fundamentally different from a full reference dictionary and even different from a comprehensive translation dictionary. The latter two must necessarily be larger than the first; however, their greater size does not ensure greater translation capability for a particular document.

ALPS produces two basic dictionaries for its interactive computer-aided translation system. One is the basic general Starter dictionary and the other is the Flagwords dictionary. The master English dictionary, which by summer 1985 will have been fully incorporated into the four bilingual English-to-other-language dictionaries (French, German, Italian and Spanish), contains between 10,000 and 11,000 source terms. The master English Flagwords dictionary contains about 1,000 terms. These are primarily function words; however, many extremely common terms have also been included such as days of the week and months.

The choice of words included in the ALPS Starter and Flagwords dictionaries is supported by analysis of the Brown University Corpus. This corpus, assembled under the direction of Dr Nelson Francis and Dr Henry Kucera, is composed of 1,000,000 words of live data representing standard American English. It contains five hundred samples, each approximately 2,000 words in length, chosen from scientific and learned literature as well as newspapers and periodicals. Frequency counts of all words appearing in the Corpus reveal the words most commonly used in American English.

Presently, the single word source terms found in ALPS master Flagwords dictionary comprise approximately 55% of the 1,000,000 words of running text represented in the Brown Corpus. The single word terms in the master Starter comprise approximately 37% of the words in the Brown Corpus.

This means that the fewer than 10,000 single word terms (in their base forms) found in these two dictionaries represent about 92% of the words found in the Brown Corpus.

In order to properly interpret the significance of this statistic, two characteristics of the Brown Corpus should be considered: (1) The Brown Corpus contains selections of a wide variety of American English but does not represent technical material except as a subgroup of the language as a whole. (2) The Brown Corpus is over 20 years old. New words have entered the language with new technologies and old ones are being used in new ways.

Because specialised vocabulary can comprise a large portion of any technical text, the first point stated represents a potential problem. However, ALPS has always maintained that Starter, Flagwords and the user's own specialised/ technical dictionaries are necessary to produce a good translation. Since it takes only a few seconds for the translator to add new words to the dictionary, development of special dictionaries suited to his or her particular needs is easily accomplished. Document dictionary tailoring is also a vital prerequisite to obtaining a good translation. The ALPS dixtraction process automatically creates document-specific dictionaries by merging relevant entries from any system dictionaries specified by the user.

The second potential problem is greatly reduced because ALPS dictionaries are so easy to update. Keeping these two points in mind, ALPS bilingual dictionaries will be able to handle about 92% of the eneral English represented by the gBrown Corpus with fewer than 10,000 single word base form terms by the summer of 1985. This extremely positive statement should be moderated with the caution that even having 92% of the "words" does not ensure that ALPS dictionaries will be able to handle word senses for all of the 92%. On the other hand, since this percentage reflects only single word terms, this says nothing about the many multiple word terms in the dictionaries and the power they give to the system.

This statistic suggests that a dictionary of 50,000 words is not inherently five times better than a carefully constructed 10,000 word dictionary. Anyone who claims this has probably included an additional 40,000 words to try to bridge the gap between what 10,000 carefully chosen terms would give and what remains. Even an additional 40,000 words may not suffice because of the highly technical nature of much of the vocabulary used by the potential customers that machine translation companies typically try to target.

Specialised vocabulary belonging to one field of technology need not and should not be mixed in with general vocabulary. Dictionaries become infinitely more manageable and less opaque when broken out into fields of specialisation. The size of the specialised dictionaries will vary from company to company, but they will contain only those terms that are peculiar to that company's technology and will not contain the tens of thousands of extraneous terms which any dictionary on the order of 50,000 terms must do.


Automatic Language Processing Systems
100 West 800 North, Provo, Utah
February 1985