Tamás Magay

TECHNICAL OR GENERAL: PROBLEMS OF VOCABULARY SELECTION IN A MEDIUM-SIZE BILINGUAL DICTIONARY

Introduction

The great sculptor Rodin was once asked how to make a statue. Take a block of stone, he said, and just carve off what is unnecessary.

How do we select vocabulary? Just leave out what is unnecessary? Is vocabulary selection really as easy as that? Dictionary-making is an endless string of decision-making. The first step is to plan what to include and what to exclude. There are 3 factors bearing on this:

(a) Size of the dictionary, ranging from the comprehensive or unabridged (100,0000 entries or more) through the concise or medium-size (between 40 and 60,000 entries) to the small or pocket dictionary (less than 20,000 entries). Two basic facts are evident: first, whatever the proposed size of the dictionary, the compiler of a general-language dictionary always has to deal with the entire lexicon of the language. Second, the more restricted the vocabulary, the greater or riskier the task of selection becomes. Critics and users are unforgiving; they will judge the quality of a dictionary first and foremost by what they do not find in it. Philip Gove was right when he stressed (1967:5) that "the function of a dictionary is to serve the person who consults it".

(b) Purpose and scope. Is the dictionary intended for reading contemporary literature, does it cover technical texts, does it serve educational aims, does it cater for the spoken colloquial language?

(c) Orientation of the dictionary. What kind of user is envisaged (cf. Al-Kasimi 1977, Harrell 1962), and what kinds of activities is it meant to serve?

Compilation and selection usually overlap. One has to pursue different policies according to the size, purpose, and orientation of the dictionary. Much of the data-gathering is in itself a selection process. If we take the English vocabulary to contain around one million words, even for the biggest available bilingual dictionary this implies leaving out some 900,000 words in the manner of Rodin.

Criteria and tests

Selection of lexical units has always been a highly subjective matter, depending first and foremost on the personal inclinations of the lexicographer, his qualifications and social outlook. Practising lexicographers will agree that whenever an old edition of a well-known dictionary comes up for revision, the compiler is possessed by an uncritical new-word-happiness, incorporating even the most short-

lived vogue words that may not stand the test of time.

And while we are aware of the fact that (1) the subjective element will always be decisive in the selection process and (2) language is an open system of an infinite number of lexical units, we must still ask: are there any objective criteria?

In selecting items from the general vocabulary, frequency counts may be of great help. However, when we are concerned with technical vocabulary and the right balance between technical and general words, I am not entirely convinced yet of the benefits of computer-aided frequency studies. The so-called general lexicon has always contained elements of the so-called specialized vocabulary, but during the last generation or so the proportion between them may have changed drastically. Today we are bombarded from all sides with technical terminology from politics, economics, electronics, medicine etc., and whether we like it or not, these fields are influencing the common word-stock of everyday language.

For the compiler of a general medium-size bilingual dictionary, these considerations require new approaches to the problem of drawing a line between general words and technical terms. In order to clarify some of these relationships, I chose just one segment of a particular field, viz. names of diseases, and tested a number of dictionaries, e.g. the LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH (LDOCE) and Országh's ENGLISH-HUNGARIAN CONCISE DICTIONARY, and word-frequency counts (Thorndike and Lorge 1944, Kučera and Francis 1967, Dahl (1979) against a list of about 400 items extracted from the Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death (henceforth referred to as the Manual). The Manual, compiled by the World Health Organization in 1977, contains some 50,000 names of diseases.

For purposes of comparison I selected List D from the Manual, a "list of 300 causes for tabulation of hospital morbidity". Out of this number, LDOCE does not include 130 (nearly half). Some of these are: avitaminosis, cleft lip, deflected nasal septum, duodenitis, helminthiasis, Hodgkin's disease, hypertension, infarction, iritis, malposition (of uterus), mycosis, nephrosis, nutritional deficiency, oophoritis, osteoporosis, otitis, prostatitis, rhinitis, uraemia, etc.

This is by no means intended as a criticism of LDOCE's selection policy, since it does in fact contain 100 more items than the Manual. I rather want to demonstrate how difficult it is to be comprehensive and consistent in finding the 'most important' (say) 400 words for diseases. Comparing Országh's ENGLISH-HUNGARIAN CONCISE DICTIONARY - considered a pioneering work especially in its selection of entries - I found approximately the same inconsistencies.

The Manual has many more items, mostly compounds, which are not entered in LDOCE as they are linguistically predictable collocations, e.g. alcoholic psychosis, bacillary dysentry, anomalies of the circulatory system, disease of the liver/nerve/skin etc., malignant neoplasm of stomach/intestines/digestive organs etc., disorders of menstruation, complications of pregnancy, internal injury, etc. While none of these collocations can be expected in the monolingual dictionary, they could or should be entered in the bilingual

dictionary according to its size, purpose and orientation.

In an L1-L2 dictionary, especially in the case where the structure of the source language is totally different from that of the target language (as English and Hungarian), many more of the diseases listed in the Manual must be included, e.g. dislocation of hip, internal injury, open wound, disease of the liver, hernia with obstruction, detachment of retina, deflected nasal septum, pulmonary embolism, delivery with complications etc., to allow for divergences in word formation, which may be more predictable in the English-foreign language part than in the foreign language-English part.

There are only a few collective nouns (such as respiratory diseases) in LDOCE which abound in the Manual. Again, such items may be predictable collocations in a monolingual dictionary; in a bilingual dictionary, however, such combinations as diseases of the urinary system or viral disease should be entered.

When LDOCE was compared with word-frequency counts such as Kučera and Francis (1967) and Dahl (1979), nearly all names of diseases missing in LDOCE could not be found in these lists either. On the other hand, some words that are firm entries in LDOCE, such as appendicitis or nephritis, are not listed in frequency counts, whereas, surprisingly enough, appendectomy and nephrectomy appear in Dahl (1979).

These tests involving just a tiny fragment of technical terms (50,000 medical terms as against the total vocabulary of 40,000 in a medium-size dictionary) suggest, first, that subjectivity is the dominant factor in vocabulary selection, second, that word-frequency counts are of little use when specialized terms are to be integrated into the general vocabulary, and third, that experiments with other technical fields such as zoological or chemical nomenclature may, mutatis mutandis, lead to the same results.

Possible approaches to the problem

It is clear that a selection policy based on some tangible principles must be put to work to support the dictionary maker's intuition, as he tries to determine the necessary and sufficient number of technical terms to be included. The lexicographer has to differentiate between what is new and what is really important, i.e. what the user really needs, and to assess the right proportion between general and specialized vocabulary.

I favour a sociolinguistic approach, since what the learner-user needs is strongly related to his social status, his level of education, and the activity for which he consults the dictionary.

Some assistance may come from continually updated terminological data banks, supported by efforts at international standardization.

For the compilation of a medium-size bilingual dictionary it is necessary to scan the widest possible sources, such as secondary-school textbooks, popular periodicals and daily papers. (I should be glad to hear whether this is being carried out systematically anywhere.) It must be borne in mind that living in a rapidly changing world involves complex education and complex vocabulary.

As Al-Kasimi has stressed (1977:31), "... in a general dictionary the vocabulary of all fields of knowledge should be represented". But the problem of how this can be achieved remains acute. Zgusta's suggestion (1971:245) that "preliminary inventories of technical terms from the single sciences" should be recorded seems good advice to follow. The first step might be to make a rank-list of all the specialized fields of human knowledge and activity. Then, within each field, the specialists themselves could help determine the relative frequency of each term. This may be done simply by asking the specialist to give a "cross-section of what the generally educated man might be expected to understand" (Pei 1966:xi). This method might then be aided or replaced by computer techniques. Some time ago Zgusta was rather pessimistic about statistical methods (Zgusta 1971:246-247). Ten years later, Makkai stressed the importance of an "updatable computer storage with every entry coded as to its ecological frequency" (in Zgusta 1980:128).

Personally, I would favour the idea of making frequency dictionaries in all the major subject fields - of the type that have been published in the Soviet Union - on the basis of international cooperation (cf. Füredi 1982). Whether traditional human methods or computeraided procedures are used, help must be given to the practising lexicographer because he is often lost in a sea of words and terms, having to make haphazard choices.

Conclusion

I have tried to demonstrate the difficulties of vocabulary selection in a medium-size bilingual dictionary with regard to the proportion of technical (specialized) and general vocabulary. The problem must be approached in a complex way and from the learner-user's point of view rather than merely linguistically. Help ought to be given to the non-specialist lexicographer; this can be expected only on an international and interlingual basis, and in a more sophisticated way than has been attempted so far. What is needed is closer contact between dictionary makers and specialists, more feedback from the learner-user, a wider context for data collection, continuous updating of data stores, further developments in specialized statistical methods and frequency counts, and closer cooperation between terminologists and lexicographers.

To put it simply: we need a new and unified effort for the benefit of the user with a complex education.

References

Al-Kasimi, A. (1977) Linguistics and Bilingual Dictionaries. Leiden: Brill

Dahl, H. (1979) Word Frequencies of Spoken American English. Essex, Connecticut: Verbatim

Füredi, M. (1982) "The usability of Russian frequency dictionaries in language teaching" Paper presented at the Győr Conference on the Computer and Language Teaching sponsored by the Committee for Applied Linguistics of the Hungarian Academy of Sciences and the College of Telecommunications, Győr

Gove, P.B. (1967) "The dictionary's function" in The Role of the Dictionary ed. by P.B. Gove. Indianapolis: Bobbs-Merrill

Harrell, R.S. (1962) "Some notes on bilingual lexicography" in Problems in Lexicography ed. by F.W. Householder and S. Saporta. Bloomington: Indiana U.P.

Kučera, H. and Francis, W.N. (1967) Computational Analysis of Present-Day American English. Providence, Rhode Island: Brown U.P.

Pei, M. (1966) Language of the Specialists. A Communication Guide to Twenty Different Fields. New York: Funk & Wagnalls/Reader's Digest

Thorndike, E.L. and Lorge, I. (1944/72) The Teacher's Word Book of 30,000 Words. New York: Columbia University Teachers College

WHO (1977) Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death. (Vol. I) Geneva: World Health Organization

Zgusta, L. (1971) Manual of Lexicography. The Hague: Mouton

Zgusta, L. ed. (1980) Theory and Method in Lexicography. Western and Non-Western Perspectives. Columbia, S.C.: Hornbeam Press