Željko Bujas

Testing the Performance of a Bilingual Dictionary on Topical Current Texts*

1.1. Of all types of reference works it is probably the bilingual dictionary that most often fails to satisfy its users. Some of the objections can, it is true, be met by saying that a dictionary — particularly of medium size (in the 40—60,000 head-word range) — cannot be expected to contain *every* word. Especially when the language in question is as lexically intensive as English. Often, however, the disappointed user will rightly point out that the foreign word (collocation or phrase), missing from the dictionary, is of frequent occurrence in the texts which he reads, or that it is very topical.

1.2. Confronted with these objectives, the lexicographer is aware of two problems involved here. One: ensuring that the really frequent items are included in such a bilingual dictionary, particularly the foreign-to-native part. Two: achieving and maintaining the up-to-dateness of this dictionary. In fact, there is a chronological link between these two problems. Problem one corresponds to the earlier and principal stage of any lexicographical venture — the selection of entries and the compilation of the dictionary core. Problem two can be stated as: how to keep up and improve the efficacity of the finished product (the bilingual dictionary) so that it does not fall behind the development of its own matter, that is of the vocabulary of the foreign language which it is supposed to interpret.

1.2.1. A simple, strict methodology offers itself for the solution of the first problem. The entries are to be selected according to frequency lists, which is particularly favorable for

---

English-to-native bilingual dictionaries in view of the available large frequency lists of English.[1] In practice, however, this is not quite so, and these lists may only serve as the skeleton for the dictionary being compiled. The bulk of dictionary matter remains to be compiled by various pragmatic procedures not be entered into here.

1.2.2. The solution to the other problem — the up-to-dateness of the bilingual (notably foreign-to-native) dictionary — lies unfortunately in the prosaic drudgery of compiling dictionary matter from current, as topical as possible, texts in the language of the dictionary in the making. Periodic publication of special "dictionaries of new words" for the foreign language involved is certainly a help in this procedure, though largely as a means of checking the material already inserted, since most neologisms from such dictionaries will as a rule be two to three years old. Carefully and systematically reading current newspapers, journals and (to a lesser degree) fiction on his own, a conscientious lexicographer will be able to register most new words and phrases immediately upon their first appearance in the texts mentioned.

1.3. Such analytical lexicographic reading of texts also serves as a direct *test of the performance of a bilingual dictionary*. A careful reading of foreign texts accompanied by systematic lookups in the bilingual dictionary tested, and by painstaking recording of every instance of the dictionary's inadequacy, is the most reliable way to compile the material absent from the dictionary. In addition to its obvious potential for completing and updating a dictionary, this material — lexicographically categorized (by "type of absence") and quantified — can yield precise numerical *performance indices* for the dictionary being tested. It is not hard to imagine a procedure which, allowing for the volume and purpose (genre) of the dictionary, would produce what may be termed *performance coefficients* for any dictionary tested in this most adequate manner: through its use on the texts for which it has been intended. By testing a dictionary on various textual genres in this way, one could (according to the highest coefficient) establish the kind of texts that the dictionary is in fact intended for. It is quite possible to imagine cases where the purpose of the bilingual dictionary, as indicated by such a coefficient, will be at variance with the very idea of the dictionary's author about

---

[1] The most recent: H. Kučera and W. N. Francis, *Computational Analysis of American English*, Brown University Press, Providence, Rhode Island, U.S.A., 1967; J. B. Carroll, P. Davies and B. Richtman, *Word Frequency Book*, Houghton Mifflin & American Heritage, New York, 1971.

the purpose of his work (as expressed, usually, in the fore-
word.) Thus, I am convinced, quite a few bilingual "general
purpose" — or even "current" — English dictionaries would,
after such testing on current topical texts, turn out to be in
fact most suitable for reading British literary texts of the
19th and the early 20th centuries.

2·1. I would now like to back up these general considerations
by presenting an actual procedure, used to test the perfor-
mance of a particular bilingual dictionary and, at the same time,
to amend, improve and update it. The volume in question is
the English-Croatian dictionary by Milan Drvodelić (1898—
1965), occupying a secure place in Croatian lexicography among
the bilingual medium-sized general-purpose dictionaries, but
also enjoying wide use in the entire Serbo-Croatian linguistic
area. Its latest edition, the fourth, was published in 1973, with
preceding ones in 1954, 1962 and 1970.[2] These four editions
(henceforth referred to as DEC 54, DEC 62 etc.) include three
revisions of the work: in 1962, 1970 and 1973. As the reviser,
I have tried very hard to make this dictionary as efficacious
a reference work as possible (within its format) for reading
current topical texts in English. This, in my interpretation,
included American and British newspapers with a very broad
spectrum of contents and styles (ranging from the seri-
ous *The Times* and *The New York Times* down to tab-
loids, weeklies (such as *Time, Newsweek, The Observer, The
Economist*) and monthly magazines (like *Reader's Digest* and
The National Geographic). The high lexcographic potential of
the papers and periodicals listed lies, naturally, in the fact that
they cover an extremely wide range of areas of human activ-
ity (each with its specific vocabulary). In addition, some of
them — notably the American *Time* — are marked by a culti-
vated updateness and inventiveness of the language they use.
Last but not least, the "current topical texts" certainly include
American and British 20th-century fiction, current pop litera-
ture (notably crime fiction and comics), satirical, avantgarde
and underground publications.

---

[2] In the course of its four editions (publisher: Školska knjiga, Za-
greb), the dictionary underwent the following changes in its title and
volume: (1954) *Englesko-hrvatski rječnik*, 976 pp. (946 without appen-
dices); (1962) *Englesko-hrvatskosrpski rječnik*, 1,104 (1,067) pp.; (1970)
Englesko-hrvatskosrpski rječnik, 1,198 (1,163) pp.; (1973) *Englesko-hrvat-
ski ili srpski rječnik*, 1,161 (1,126) pp. — owing to denser print.
    These changes in title reflect changes in official attitudes, in no
way affecting its contents. Thus, the native glosses are exclusively in
standard Croatian, nor could they be anything else in view of the
ethnic origins and cultural-educational backgrounds of both the author
and the reviser.

2.2. After my first two revisions (in 1962 and 1970), I considered the DEC ready to be thoroughly tested, and tested on the very type of texts for the reading of which it had primarily been expanded and updated. So, over two academic years (1971/72 and 1972/73), a total of 18 analysts[3] under my guidance tested the efficacity of the DEC in its (then) latest edition of 1970.

3.1. Carefully reading the listed and similar publications in American or British English,[4] the analysts, it turned out, had followed two procedures, their choice depending on individual temperament. Some relied upon intuition, looking up in the DEC 70 only those words (expressions, phrases) which they expected to be absent from the dictionary. This group, naturally, needed more text and some of them went through as many as three issues of a newspaper or magazine before compiling the material assigned (400 items as a rule). The other group looked up practically every word in the text analyzed (except for the "functional items": the articles, conjunctions, pronouns, prepositions, etc.), filling their quota after only 30 to 50 pages of text. Analysts from this group found more items where, in their view, the dictionary offered an inadequate Croatian translation. They also discovered a few serious omissions which had occurred in previous editions. Thus, they noticed that the headwords *abbey* and — hard to believe — *role* were absent from the DEC 70. But while these two were already missing from the first (1954) edition, *trout,* present in that edition, mysteriously disappeared from the second and third editions. Thanks to such rigorous checking and non--reliance on intuition, the DEC 73 now contains both *abbey* and *role* and *trout.*

3.2.1. Using one or other of the techniques described, the analysts, when they came across an instance of inadequate performance by the GEC 70, would first make out a separate slip for any such "item of inadequacy". Every such item, under-

---

[3] All of them undergraduates in the English Department, Filozofski Fakultet, University of Zagreb, who did their analyses in the form of English Language Graduation Theses. Their names are: Eda Barbalić, Nevenka Blažević, Drenka Drezga, Višnja Butraković-Dravinski, Svjetlana Dupalo, Snježana Jakšić, Darija Kosić, Marijana Leskovar, Maja Pribanić, Zdenka Pula, Karmen Relja, Lelija Sakač, Neda Šegota, Dubravka Šnajder, Ljerka Štibuhar, Marina Šulc, Dunja Uvodić and Katerina Verbanac.

[4] A total of 34 issues were analyzed: *Newsweek* (8 issues), *Time* (7), *Reader's Digest* (7), *the Economist* (4), *the National Geographic* (3), *the Daily Mirror* (2), *the Sunday Express* (1), *The Times* (1) and *the Herald Tribune* (1). Dates of publication: January 1970 through August 1973.

lined, would be written out with the rest of the sentence as a context. The title of the publication (with date and page) and the Croatian translation of the item underlined, supplied by the analyst, were also recorded on the slip.

3.2.2. After collecting the required number of items (slips), the analysts would go on to the next stage — the classification of the material collected into one of three basic "types of inadequacy":

> Type 1. The item underlined is completely absent as a headword from the DEC 70.
>
> Type 2. The item underlined is present in the DEC 70 as a headword, but absent in the particular collocation (phrase).
>
> Type 3. The item underlined is present in the DEC 70, but its Croatian translation (gloss) in the dictionary is inadequate.

3.2.3. After this, in the third stage of the procedure, the analyst, assuming the role of an active lexicographer, would break down the three basic types into two categories. One was for the items recommended for lexicographic intervention, that is for inserting them into the DEC 70 or otherwise correcting the inadequacy observed. The other category covered the items for which no such recommendation was made. In either case, the analyst would specify the lexicographic reasons for his decision, thus creating a number of subcategories. Here is a full survey of these (sub)categories illustrated by concise examples.

> **(1) Insertion recommended** because the item underlined:
>
> a) is common /1a/
> (type 1) *weaponry:* naoružanje
> (type 2) *fun fair:* luna park
> (type 3) [*touchdown*]: spuštanje, ateriranje
>
> b) is topical or a neologism /1b/
> (1/1b) *spacecraft:* svemirska letjelica
> (2/1b) *punched card:* bušena kartica
> (3/1b) [*hijacker*]: otmičar aviona
>
> c) requires minimal space /1c/
> (1/1c) *publicly:* javno
> (2/1c) [*sewer*] (i ~ *system):* kanalizacija
> (3/1c) [*pandemonium*]: urnebes
>
> d) is something typical of British or American society /1d/
> (1/1d) cowpoke: kauboj

      (2/1d) *National Guord: američke teritorijalne jedinice*
      (3/1d) *[headdress]:* (indijanska) perjanica

(2) Insertion not recommended because the item underlined:

  a) is not obligatory for a medium-sized dictionary /2a/
      (1/2a) *jacaranda:* brazilsko tropsko drvo
      (2/2a) *optional extras:* neobavezan dodatni pribor
      (3/2a) *[put (A com)]:* kupnja

  b) is too technical /2b/
      (1/2b) *muom:* pozitivno ili negativno nabijen meson
      (2/2b) *stereo phono cartridge:* stereofonska kaseta
      (3/2b) *[point]:* klema

  c) is clear by itself /2c/
      (1/2c) *antiburglary:* protuprovalni
      (2/2c) *peace lover:* ljubitelj mira
      (3/2c) —

  d) requires context or descriptive translation /2d/
      (1/2d) *patch:* greben[5]
      (2/2d) *car pound:* policijski parking za vozila koja su napravila prometni prekršaj
      (3/2d) —

  e) is a pun /2e/
      (1/2e) —
      (2/2e) *curry on (=carry on):* nastaviti živjeti u Indiji[6]
      (3/2e) —

3.2.3.1. As we can see, of the 27 theoretically possible categories (3 basic types, inserted or not for 9 listed reasons) only four were not encountered in the course of actual testing. The absence of examples for categories 3/2c and 3/2d is logical if we take a closer look (Croatian glosses are directly linked to the English entry in the relation of a translation to its original). Instances of puns in Croatian glosses (3/2c) are plainly impossible. Finally, the absence of category 1/2c cases is simply the consequence of an insufficiently large corpus used to test the DEC 70.

3.3. The precise numerical distribution, by type and category, of all the vocabulary item recorded in testing the DEC 70 (totaling 6,272) is presented in the following table:

---

    [5] Full context: *Mistaking distant coral patches for white marble, he planned to build a city there.* (Reader's Digest, December 1972, p. 81).

    [6] In the sentence: *Although he lost his official post when India became independent, he decided to curry on.* (Reader's Digest, December 1972, p. 86)

Table 1

Distribution of "Items of Inadequacy" in DEC 70

Reason for (non)inclusion	(1) No English entry	(2) No English collocation	(3) Croatian translation inadequate	TOTAL
1 a) common item	1,441	1,069	725	3,235
b) topical item	240	126	125	491
c) minimum space	12	52	934	998
d) typical of AS societies	113	56	15	184
Total included	*1,806*	*1,303*	*1,799*	*4,908*
2 a) not for dictionary of this size	244	112	37	393
b) too technical	132	76	8	216
c) clear by itself	321	416	—	393
d) requires context or descriptive translation	4	13	—	17
e) pun	—	1	—	1
Total non-included	*701*	*618*	*45*	*1,364*
GRAND TOTAL	2,507	1,921	1,844	6,272

3.3.1 Of the many conclusions made possible by this table, only the most important will be pointed out and commented on:

1. Of the 6,272 items of inadequacy encountered in the DEC 70, a total of 4,908 (or as many as 78.3%) have been recommended for insertion. It should be remembered, however, that this total includes all the repetitions of the same item which are unavoidable with 18 analysts testing the dictionary independently of one another. In addition, one-fifth of all examples in this group come from the "minimum of space" category which further reduces their mass.

2. Type 1 is dominated by examples from the category "common item" (totaling 1,441 or 44.5%). This was to be expected from analysts who are not active lexi-

cographes. Thus, category 3/1a is somewhat under-
-represented (only 22.4% of all items), though its
less obvious potential is no inferior to that of
categories 1/1a or 2/1a.

3. Ranked by frequency, the top five individual catego-
ries are:

1a (to be included: common item)	3,235 (51.6%)
1c (to be included: minimum space)	908 (15.0%)
2c (not to be included: clear by itself)	737 (11.8%)
1b (to be included: topical item)	491 (7.8%)
2a (not to be included: size of dictionary	393 (6.3%)

Consequently, these five categories account for a total
of 5,854, or no less than 93.3% of all inadequacy items
recorded. The top-rank positions of these categories
come as no surprise since they result from the very
nature of the material and the analytical procedures
applied.

4. Several markedly lopsided ratios among the basic
types of inadequacy within one category are readily
observable. The reasons for all of them, however, are
logical. It is clear, for instance, why subcategory 3/1c
fills its category (1c) almost by itself (with 93.6%):
adding one word is largely called for in Croatian glosses.
The absence of type 3 in category 2c has already
been explained (cf. 3.2.3.1.). The low share of the
same type in 2a is easy to understand: the volume of a
bilingual dictionary is usually controlled by interven-
tions in its left-hand side. Finally, the under-represen-
tation of type 3 in categories 1d and 2b is a natural
consequence of the fact that the more important pro-
cedures for processing specialized vocabulary are simi-
larly those on the left side of a bilingual dictionary.

4.1. Another important parameter of vocabulary structure —
and consequently of the structure of a dictionary as formalized
vocabulary — is the distribution of vocabulary items among
the vocabulary areas. In our case, an investigation of the vo-
cabulary area ratios in the items compiled (observing at the
same time the subdivision into the three basic types of
inadequacy) is certain to offer insight into further important
aspects of the structure of the dictionary tested (DEC 70). This
is borne out by the figures in Table 2, made possible by an addi-
tional procedure used by the analysts in the final stage of
processing the recorded instances of inadequacy in the DEC

70. Each slip was marked with an area label standard in lexicography (such as *mil, ent, mar* and the like), with a number of pragmatic descriptive designations (like *occupations, fashions, domestic,* etc.) added. Usage labels *coll, sl, tab* — even *fig* — were also applied in an effort to reduce the number of items remaining under the final, catchall area label *gen (general),* to be read as "cannot be assigned to any specific vocabulary or usage area". The degree to which the DEC 70 analysts have been successful is shown by Table 2, with vocabulary areas ranked by descending frequency down to the total of 50 items.

Table 2

Rank List of DEC 70 Items of Inadequacy by Vocabulary Area

Type:	1	2	3	Total	% of all items	including:
gen (fig)	159	194	171	524	8.4	
gen	895	598	1,391	2,884	45.0	
government	210	211	85	506	8.1	*pol, par, dipl*
gen (coll)	77	113	72	262	4.2	
fashions	155	34	30	219	3.5	*text, cosmetics*
med	65	69	46	180	2.9	*anat, pharm, sur*
mil	82	58	26	166	2.6	*nav*
eng	70	56	33	153	2.4	*technol*
sp	40	45	20	105	1.7	
gen (sl)	47	32	25	104	1.7	*hobbies*
occupations	40	48	14	102	1.6	
jur	30	49	16	95	1.5	
chem	70	16	5	91	1.5	
mot	22	41	19	82	1.3	
domestic	36	29	13	78	1.2	
zoo	49	4	9	62	1.0	*ornith, ichth, ent*
aero	38	10	12	60	1.0	
cul	24	18	12	54	0.9	*drinks*

4.2. Out of a total of 59 vocabulary areas distinguished at this stage of analysis, the above list shows only 18, or less than a third. These 18 areas, however, cover 5,732 items or 91.4% of the total material. The non-listed 41 areas, with their 540 items, cover only 8.6% of the total DEC 70 material. The following conclusions, drawn from the Table 2 figures, are therefore highly reliable as general statements about this material.

1. The total share of all general *(gen)* vocabulary areas, with 3,781 items[8] or 60.3% of the material collected, leaves 2,451 items with 39.7% of total material to specific vocabulary areas. This is, no doubt, considerably above the average total share of this second group in general-purpose bilingual dictionaries. This comes as no surprise since one of the characteristics of topical texts is the marked presence of items belonging to specific vocabulary areas.

2. The low share of slang items (1.7%) demonstrates the efficacity of the DEC 70 in that area, because the texts on which the dictionary was tested as a rule make free use of slang. Some indeed, like *Time Magazine*, deliberately use it as a major ingredient of their distinctive style.

3. The easily predominant single category in the list is *gen*/3, accounting with its 1,391 items for almost one-quarter (22.2%) of all the material compiled. It testifies to the lexicographic sensibility of the analysts who suggested additional or more adequate Croatian glosses for items from the general area of vocabulary, with its more pronounced synonymic potential than that of the specific, technical vocabulary areas.

4. The high positions of areas *government* and *mil* on the list is convincing testimony to the primary thematic preoccupations of such contemporary texts as were analyzed in the project described here. At the same time, this is a general warning to lexicographers to be as liberal as possible with items from these vocabulary areas when updating the left side of any dictionary.

5. The equally high ranking of *med* and *eng* warns us of the permanent topical character of these areas. We must, however, bear in mind that the DEC is after all a general-purpose, not a technical, dictionary. The same is true of the slightly lower-ranking *jur*, *chem* and *aero*.

6. The high position of *fashions*, also of *sp*, *occupations* and *cul* (but not of *mot!*) comes as a surprise. Though we must indeed remember the space restrictions of the DEC, this probably is a consequence of the dictionary's inadequate performance in these vocabulary areas. Thus,

---

[7] Of which 3,774 are in the table — only 7 items coming from lower frequencies. In fact, each of the 7 represented one area: *gen (derog)*, *gen (dial)*, *gen (hum)*, *gen (obs)*, *gen (poet)*, *gen (rhet)* and *gen (tab)*.

it is something that should be kept in mind when later editions are prepared.

5.1. In conclusion, I would like to illustrate, by means of a short excerpt from the DEC 73, the volume and usefulness of all the additions and amendments included in that, the fourth, edition of the dictionary, after their extraction from the collected material described in this paper. Here then is the full list of all items beginning in BA (roughly one-fifth of the letter B) included in the dictionary from this material. The category and (in parentheses) frequency of occurrence in the material have been added as additional parameters which can, and should, influence the decision about inclusion.[8] The brackets have been used to indicate the portion of the item already present in the DEC 70.

2/1a	(1)	*back away (off)* uzmaknuti, povući se
1/1a	(6)	*backdrop* prospekt (kulisa) *(theat);* pozadina *(fig)*
1/1b	(5)	*backlash* reakcija na; protuakcija
1/1a	(2)	*backstage* iza pozornice
1/1a	(2)	*backtrack* reterirati
1/1a	(3)	*back-up* podrška
1/1a	(1)	*backyard* stražnje dvorište *(Am)*
1/1a	(1)	*bafflement* zbunjenost
3/1c	(1)	*[balloon]* naglo porasti *(fig)*
3/1c	(1)	*[bankroll]* kapital *(fig)*
3/1b	(1)	*[banner]* transparent
1/1d	(1)	*bar mitzvah* židovska proslava zrelosti (13-god. dječaci)
1/1d	(1)	*barracuda* uc, škaram (grabežljiva riba)
3/1c	(1)	*[barrage]* uragan *(fig)*
1/1a	(1)	*barrel-chested* snažnog (izbočenog) grudnog koša
3/1d	(1)	*[base¹]* „etapa" u bezbolu (ugao glavnog polja) *(Am sp)*
3/1a	(1)	*[base²](~d in)* sa sjedištem u
3/1a	(4)	*[batch]* serija
1/1a	(1)	*bath-cube* (mirisna) kocka za kupanje
3/1a	(1)	*[baton]* dirigentski štapić, palica
1/1d	(1)	*batter* branilac palicom (u bezbolu) *(Am sp)*
3/1c	(1)	*[battery]* niz *(fig)*
3/1a	(1)	*[batting]* sloj vate (vune)

5.2. The letter B was selected as an average representative of the non-Latin and non-Greek part of the English vocabulary

---

[8] As each analyst made out only one slip for each item (on first encountering it as an instance of DEC 70 inadequacy), these figures have word-type rather than word-token significance (cf. Kučera & Francis, op. cit., pp xx and xxi). Naturally, this does not lessen their value as dispersion parameters.

(unlike, for instance, the headwords in A, C, D), where the high density of abstract, rare and technological entries calls for a somewhat different approach to the selection of lexicographic material. As the above-listed 23 items are spread over 12 pages of the DEC 73, we can roughly estimate the average density of included items for the entire dictionary to be 2 items per page. This works out at some 2,200 items for the 1,126 pages of the DEC 73. Of these, about one-half will probably belong to basic category 1 (in keeping with the 12 : 23 ratio observed in the above list), fully representing new headwords. The total of some 2,300 items actually included in the DEC 73 from the mass of 4,908 items recommended for inclusion (cf. Table 1), is a realistic ratio in view of the already mentioned unavoidable repetition of the same items among the 18 analysts. (The above list itself, when frequencies of occurrence — i. e. repetitions — are taken into consideration, represents a total of 39 occurrences of the items listed.)

6.1. In conclusion, looking back to the material collected and some insights made possible by it, I would like to point out that such dictionary-performance analyses are valuble in two respects:

6.1.1. No words need be wasted, I am sure, to prove their usefulness in practice. A mere glance at the short list presented above of additions to a bilingual dictionary, convincingly demonstrates their value and the obvious topicality of practically every bit of the material included, following methodology that has been described in this paper. The full implications of inclusion of 2,200 such items for modernizing and generally improving a medium-sized dictionary will be best appreciated by active lexicographers.

6.1.2. A certain contribution of this paper on a higher, theoretical level of linguistic research may be less obvious. Still, I will venture the statement that this paper has indicated some intriguing quantitative correlations among vocabulary areas in general (and specifically in terms of topicality). And this is certainly one of the aspects of dictionary makeup as a formal expression of lexical structure. Similarly, the introduction of the term *performance coefficient* in bilingual lexicography (while pointing to one of its less obvious possible applications) means an expansion of the scope of quantitative methods in lexicography. Both these contributions also have a more general linguistic value, since they are attempts at a certain formalization within lexis, a sector of linguistics calling for much additional effort.

204