Leonard NEWMARK, University of California, San Diego

Reversing a One-Way Bilingual Dictionary

Abstract

In 1996, I completed a 10-year project to compile an Albanian-English dictionary of some 75,000 entries and sub-entries. With that project completed, I decided to prepare a companion English-Albanian dictionary, but I did not want to devote another 10 years to that compilation. Instead, I wanted to see how far purely computational techniques would go in converting my dictionary into its derived reverse. This paper is a report on the degree to which the attempt succeeded and the degree to which human intervention was required. Examples are provided to illustrate some rather surprising results, and a general conclusion is drawn for bilingual lexicography.

For the past year I have been trying to convert a computerized bilingual dictionary automatically into another dictionary in the reverse direction: specifically my Albanian-English dictionary, designed for readers whose access language is English, into an English-Albanian dictionary for writers. My source dictionary, published in March 1998 by the Oxford University Press as *Oxford Albanian-English Dictionary*,[1] has some 75,000 entries and sub-entries, more than any other dictionary of Albanian, bilingual or monolingual. It boasts a number of features that distinguish it from many other bilingual dictionaries: 1) inclusion of large numbers of non-standard items (indicated by the symbol •) as well as all attested standard stems; 2) marking of morpheme boundaries in Albanian words; 3) inclusion of some 16,700 phrasal expressions, in particular, phrasal names, collocations, idioms, and proverbs; 4) use of large numbers of bipartite definitions with a discursive description of the sense followed, after a colon, by English synonyms exemplifying that sense; 5) inclusion of large numbers of terms for grasses, flowers, birds, and fish with their scientific definitions; 6) inclusion of a modest amount of encyclopedic information to explain words whose strictly lexical meaning would not make their use in Albanian contexts intelligible; 7) listing of the various stem forms of lexemes as separate entries in their own alphabetical position to enable readers to decipher otherwise mystifying forms encountered in actual texts; 8) indication of the specific limits of variation of phrasal expressions that leave idiomatic senses intact (e.g., by using the citation forms of variable verbs in such expressions and marking them with a symbol • at the end of the stem); 9) reverse-alphabetic listing of the potential grammatical interpretations of all possible word endings in an appendix, to permit readers to figure out puzzling words by working from back to front; 10) omission of examples for most senses, in exchange for the space and time needed to provide more head entries;[2] 11) elaborate labeling of Albanian domain and register distinctions; 12) rendition of phrasal expressions by stylistically similar English expressions, frequently supplemented by literal translations (between quotation marks) to enable more nuanced understanding.

The computer files from which the Albanian-English dictionary was generated are plain text files embedded with simple visible 2-letter codes (e.g., HW [headword], DF [definition], TK [technical name], CO [collocation]) immediately preceded by a period (.) and immediately followed by the part of each entry that is to get that formatting. The easily redefinable codes are later translated by a set of UNIX scripts into formatting instructions in TeX, which can go directly to a printer or indirectly by translation into PostScript files. The simplicity of such

transparent and flexible coding, in contrast with elaborate schemes requiring complex coding by experts into predefined structures,[3] was initially dictated by limits typical of languages that attract little commercial interest and thus little financial support. The formatting codes, together with the commas (that separated equal alternates) and semicolons (that separated sub-senses) in the definitional portion of the head entries, served as the crucial boundary markers in the search and replace operations that created the first draft of the reversed dictionary.

At the beginning, I did not really know what to expect from an automatic reversal process. Since codes in the computerized base of the source dictionary identify the main parts of the entries, I knew that there would be no problem with single-word equivalents like

collegial *adj* kolegjia'l

particularly when the equivalents are of the same register in both languages:

collegiality *nm (Bookish)* kolegjial|ite't

Entries with strings of single-word equivalents for the Albanian word could be converted automatically by duplicating the entry for each equivalent, while moving that equivalent into head position. Thus, the entry

skër|fy'ell *nm [Anat]* trachea, windpipe

could automatically be duplicated, reversed, and then included under both the head entries **trachea** and **windpipe** as identical sub-entries: **trachea, windpipe** *nm [Anat]* skër|fy'*ell*

I was surprised and happy to discover that such sub-entries incidentally (and accidentally) offer useful information that is not apparent in bilingual dictionaries compiled in more traditional ways: the Albanian equivalent serves neutrally both as a specialized word **trachea** and as the ordinary word **windpipe**.

It seemed unlikely, even absurd, however, to expect success for entries with more complex multi-word definitions. And indeed explanatory entries like

my'kë *nf* blunt side of a bladed tool

defied automatic conversion, since none of the component words of the English explanation could be translated as **my'kë**. However, emboldened by the success of the conversion of the strings of single-word equivalents, I tried a modification. For these entries, I first marked one or more key words in the explanation — in a few cases I had to add the key words — and then placed the reversed entry automatically as a sub-entry under that key word. Thus, under the head entry **blunt**, there is a sub-entry:

blunt side of a bladed tool *nf* my'kë

At this point, my new dictionary began to look like a dictionary thesaurus, not an unwelcome development for a language like Albanian for which no other thesaurus exists, nor is likely to exist otherwise. I will return to this point later.

The next set of at-first-sight intransigent entries consisted of the thousands of phrasal entries in my Albanian-English dictionary. Here again, however, the key-word technique produced the interesting and useful result of greatly enriching this bilingual dictionary: 1) in many cases, it provided examples and context for English and Albanian matching; and 2) it automatically added a fairly large number of phrasal entries that are missing from any other English-Albanian dictionary. Thus, the automatically generated entry **jinx**[4] now appears with an automatically generated, rich set of sub-entries:

jinx n *(Colloq)* ters

1. **jinx who attracts bad luck by mentioning it** *nm* gojë|te'rs
2. **a terrible jinx** *(Crude)* Tersi i Pojanit
3. **to be a jinx** e ka• këmbën të prapë *("to have one's leg wrong")*
4. **to be afflicted by the evil eye: be bewitched, have a jinx** *vpr* për|sy'|sh•*et*
5. **to rid [] of the evil eye: break the jinx on [], release [] from a curse** *vt* [] çl|më|sy'|sh•
6. **to jinx <>, bring bad luck to <> <>** vë'• çullin

These examples illustrate some of the consequences of reversing the Albanian-English dictionary by automatic or semiautomatic procedures. The innovative characteristics of that dictionary and consequent value to the user of the reverse dictionary would likely be lost if the dictionary had to be compiled from scratch, in large part because that value would not repay the labor necessary to include them, particularly for a language like Albanian with limited commercial appeal. The richness of phrasal entries for **jinx** is a natural consequence of the way in which the entries were prepared automatically from the source dictionary.

- The head entry for **jinx** includes a symbol *n* specially defined as implying that the Albanian **ters** is a noun stem that is masculine in this form, but having a corresponding feminine form **te'rse**. The entry includes the register label *(Colloq)*, no innovation in itself, of course; the important point is that the elaborate number of informative labels and their frequency of use in the Albanian-English dictionary is captured for free in this English-Albanian derivative.
- Sub-entry 1. illustrates the use of internal morpheme boundaries in Albanian words by the symbol | ; together with the indication of primary stress by the symbol ' after a vowel letter, such indication helps the user to pronounce the Albanian word as well as to understand its structure — in this example showing that **gojë|te'rs** is a compound whose second element is **ters**.
- Sub-entry 2. shows that vulgar expressions, marked as *(Crude)*, are included in this dictionary because they appear in the mother dictionary and are generated automatically and for free; if I were compiling the dictionary from scratch, I would not have sought out such an entry, since the corresponding English expression is of no particular value and would not have called for translation.
- Sub-entry 3. illustrates the use of the symbol • at the end of a verb stem to indicate that the verb (**ka•** 'to have' in this example) may appear in any of its inflected forms to render corresponding variation in the English form — indicated by the English infinitive: so English "Nexhmija was a jinx." may be rendered in Albanian by **Nexhmija e kishte këmbën të prapë**. This entry also illustrates the frequent use of supplementary literal translations in phrasal expressions, characteristic of the mother dictionary, but not characteristic of many bilingual dictionaries compiled in traditional ways.
- Sub-entry 4. illustrates a frequent form of definition in the mother dictionary: a bipartite definition with a discursive description followed by one or more approximate English equivalents. The grammatical label *vpr* (passive/reflexive verb) tells the user how the verb is to be conjugated; the location of • in the form **për|sy'|sh•***et* shows where the verb stem ends, and the italics after • shows that the *e* (marking *vpr* forms) and *t* (marking 3rd person singular) are evanescent, in the sense that they do not

appear throughout the conjugation. In the form **my'kë** cited above, we saw a similar use of this typographical convention to indicate that the *ë* is evanescent: it does not appear in the nominative definite form **my'ka**.

- Sub-entry 5. is again in the form of discursive description plus English equivalents. It illustrates how [] in the English expression stands for a pronoun in the accusative case, which in turn may stand for any accusative object. The placement of the corresponding [] in the Albanian construction tells the user where a corresponding accusative pronominal clitic will be if the *vt* (transitive verb) in Albanian[5] is to sound natural.
- Sub-entry 6. connects the lemma item **jinx**, using a comma, to a synonymous, somewhat less colloquial expression in English, indicating the tonal range of the corresponding Albanian expression. It also illustrates how <> in the English expression stands for a pronoun in the dative case, which in turn may stand for any dative object. The placement of the corresponding <> in the Albanian construction reminds the user where a corresponding dative pronominal clitic will be if the verb is to sound natural in Albanian.

In addition to these valuable features of the source dictionary, the semiautomatically reversed dictionary offers — with little cost in additional labor — some two thousand names of birds, flowers, grasses, fish, diseases, etc., followed by their Latinate scientific names.

shark

1. **shark** *[Ichth]* qen deti *("sea dog")*
2. **angel shark** *nf [Ichth]* skadhi'në *Squatina squatina*
3. **angular rough shark** *nm [Ichth]* peshk|a|qen-de'rr *Oxynotus centrina*
4. **bramble shark** *nm [Ichth]* yll|za'k *Echinorhinus brucus*
5. **gray shark** *nm [Ichth]* peshk|a|qe'n *Galeorhinus galeus*
6. **great blue shark** *nm [Ichth]* peshk|a|qe'n *Prionace glauca*

A major value of the derived dictionary lies in the way in which so many of its entries and sub-entries automatically provide contexts (in English) for choosing among alternative Albanian translations of an English term.

windpipe

1. **windpipe, larynx** *nm (Colloq)* gurma'z
2. **windpipe, throat** *nm (np ~j) (Colloq)* gu'rgu*ll*
3. **windpipe, trachea** *nm (in phrasal expressions)* rryl
4. **trachea, windpipe** *nm [Anat]* gabzhe'rr
5. **trachea, windpipe** *nm [Anat]* gërl|a'c
6. **trachea, windpipe** *nm [Anat]* skër|fy'e*ll*
7. **larynx, windpipe** *nm [Anat]* lari'ng

Contrast the value to a writer of an entry like this with the value of the entry for **windpipe** given in the two presently best English-Albanian dictionaries:

windpipe (uind'pajp) *n.*, skërfyell, rryl, gabzherr[6]

windpipe ('windpaip) *n.* kanali i frymëmarrjes, gabzherr, laring[7]

Of course, these two comparison definitions were written for Albanian users, rather than for English-access users; but the point of the comparison remains: the amount of information contained in the automatically generated definitions far exceeds the amount contained in either or both of the hand-generated definitions. Even Albanian users, as well as the English-

access users for whom the dictionary was originally aimed, will find value in the sheer wealth of information it offers.

While the automatic and semiautomatic compilation of entries does save an enormous amount of human energy and has some valuable consequences, there remain many editing tasks for the human editor. Human judgment is still needed to decide whether particular apparent redundancies should be expunged or left in; for example, should the three mechanically generated sub-entries

trachea, windpipe *nm [Anat]* gabzhe'rr
trachea, windpipe *nm [Anat]* gërl|a'c
trachea, windpipe *nm [Anat]* skër|fy'e*ll*

be allowed to stand, should they be coalesced, or should one or more be eliminated as merely confusing to the user? For items that happen not to have been generated at all by the computational methods, human judgment is again needed to find those items and to decide which are important enough to be added. For example, no entry for **abattoir** nor for **abbreviate** was generated from the source dictionary. Their absence I discover by comparison with other dictionaries; choosing which dictionaries to consult also requires human judgment. After judging that the first word is dispensable, but the second is not, I still need to compose the new definition by hand, with the help of an Albanian coadjutor.

As may be apparent from the few examples in this paper, the semiautomatically generated reverse dictionary ignores the boundaries between ordinary dictionary, thesaurus, dictionary of collocations, lexicon of idioms, technical dictionary, and encyclopedia — as did its source dictionary. Judging from the result, this overstepping of bounds constitutes a strength, rather than a weakness; from the point of view of the user, the dictionary gains in value by including a miscellany of information related to its words, whether lexicographic purists like it or not. Especially for a language like Albanian, for which commercial and scholarly interest is so limited, specifically focused reference materials requiring high levels of expenditure of time/money are unlikely to be developed. This paper has attempted to show a way in which a work designed for one purpose can be exploited to produce another one that serves a very different purpose, in an economical and unusually valuable way.

Notes

[1] Written with the invaluable assistance of my coadjutor, Dr. Vladimir Dervishi.

[2] I would argue that this dictionary's intended users will be readers trying to interpret problematic words in given contexts, rather than writers needing models to follow. The time and space economies within which lexicographers work require that decisions of this kind be made: in order to produce a dictionary useful to others and to get it out in a reasonable amount of time, we must all limit the amount of information it would be possible for us to ferret out and provide if we had unlimited resources. Different lexicographers make different judgments as to which and how much information will be optimal for the purposes of a given dictionary. There can be no "maximally" useful dictionary, since the amount of potential information is essentially boundless.

[3] For example, like those used in the architecture described by Willy Martin and Anne Tamm in "OMBI: An editor for constructing reversible lexical databases", *Euralex '96 Proceedings I-II*, pp. 675-687.

[4] No entry for **jinx** appears at all in any existing English-Albanian dictionary. Here is the list of those dictionaries (in descending order of comprehensiveness).

Gasper Kiçi and Hysni Aliko, *English-Albanian Dictionary*, Second Edition, privately printed in Italy, 1991 (27,000 headwords, 627 pages), offers more phrasal definitions than any of the others.

Ilo Stefanllari, *Fjalor Anglisht-Shqip*, Tirana: Shtëpia Botuese 8 Nëntori, 1986 (442 pages) is widely used in Albania.

Stuart E. Mann, *An English-Albanian Dictionary*, Cambridge: Cambridge University Press, 1957 (21,000 headwords, 434 pages), is especially rich in bird and plant names. Its central Gheg dialect forms of Albanian are considered non-standard in present-day Albania, and its stoutly British focus makes many of its entries unintelligible to present-day American users.

Ministria e Arësimit dhe e Kulturës, *Fjalor anglisht-shqip*, Tirana: Mihal Duri, 1966 (reprinted in Prishtina, Yugoslavia in 1969 and 1972 with the title *Fjalor Anglisht-Shqip për Shkolla e Mesme*) (20,000 headwords, 340 pages) is inferior to the Stefanllari dictionary.

C. A. Chekrezi, *English-Albanian Dictionary*, Boston: Ilia Chapullari, 1923 (13,000 headwords, 187 pages), in a tiny format, has enjoyed considerable popularity among older Albanian-American users .

Nelo Drizari, *Albanian-English and English-Albanian Dictionary*, New York: Frederick Ungar Publishing Co., 1957 (184 pages in English-Albanian section), is full of inaccuracies.

Ylvi Basha and Muhamet Kapllani, *Fjalor themelor anglisht-shqip*, Tirana: Shtëpia Botuese e Librit Shkollor, 1972 (2500 headwords, 74 pages).

I. Stefanllari & V. Dheri, *Fjalor frazeologjik Anglisht-Shqip*, Tirana: Shtëpia Botuese 8 Nëntori, 1980 (389 pages) is a hodgepodge collection of old literary quotations ("to give somebody a Roland for an Oliver") and modern colloquial expressions ("to give somebody credit for"), without distinguishing the two and with Albanian glosses not usable as equivalents.

Sadik Berisha, *Fjalor Praktik Drejtshkrimor dhe Drejshqiptimor Anglisht-Shqip*, Prishtinë: Enti i Teksteve dhe i Mjeteve Mësimore i Krahinës Socialiste Autonome të Kosovës, 1984 (10,000 words, 239 pages) is just a list of English words without definitions, to be used as a speller.

[5] The rules governing the presence of pronominal clitics in Albanian are quite complex. When they do appear, however, they precede the verb, except optionally when the verb is in the imperative mood.

[6] Kiçi and Aliko, *op.cit.*

[7] Stefanllari, *op.cit.*