G. O. Tarnawsky

Tagging Techniques for Incorporating Microglossaries in an Automatic Dictionary

Microglossaries have been frequently suggested as a practical solution to some of the semantic and homonym problems encountered in automatic translation.[1-4] This Communication indicates a systematic approach to solving those problems.

The commonly proposed microglossary would contain either all the entries necessary for translating within a given technical field, or only those which may have a specific meaning in that field.

In the first case the microglossary would be the only dictionary used in the automatic translating system and would act essentially as a special-field dictionary yielding a translation tailored for that field. This means that the microglossary would have to contain about 100,000 entries and that it would probably have to be selected manually, since it would be difficult for a computer to house more than one such dictionary at a time.

The second approach would use in addition to a microglossary a general-purpose dictionary in which there would be stored those words that are common to all the fields and have the same meaning in each of them. The input text, then, would be first looked up in the microglossary for the particular field (which could be selected automatically), and only if a word was not found there would the search be directed to the general dictionary. As a consequence, the words common to all the fields would be found only during the second dictionary search. This would reduce the over-all speed of translation, especially since the most frequently used words would usually be common to all the fields.

The microglossary approach described in this Communication combines the advantages of both of the systems: it permits the microglossaries to be intermeshed in a single, general-purpose dictionary and allows them to be selected automatically by the computer. The approach is currently used in the Russian-English automatic translation system* developed by IBM. A few sample entries in the general-purpose dictionary are shown in Table 1, and the sequence of operations is shown in Fig. 1.

The virtue of the system lies in its simplicity and the ease with which it can be used. If it is desired to translate a text with the aid of a microglossary, the text is preceded by a tag specifying the pertinent field. The tag can be in English, such as *Mathematics* or *Electrical Engineering*, or in Russian, in which case it could be the title of the journal in which the text occurs, e.g., *Uspekhi Matematicheskikh Nauk* or *Zhurnal Neorganicheskoy Khimii*. This, in effect, permits the microglossary field to be assigned by the input typist. If no tag exists, it is assumed that the text does not belong to any special field, and those words whose meaning differs in various fields will be given a general translation.

As a first step in the translation process, the tag preceding the text is looked up in the dictionary, identified as such, and a unique code is stored in a fixed location in the computer memory for future reference. If no tag is found, the cell in the memory will contain a special code denoting this.

The automatic dictionary contains all the forms necessary for the input words to be found, regardless of whether a word belongs to the microglossary category or not. Words common to all the fields and having the same meaning in all of them, as well as those which may be considered to belong to a special field but which are unambiguous (e.g., *stethoscope*, *fricative*, *chromosome*), have listed with them both a translation and the information necessary for syntactic analysis. Those words whose meaning differs in various fields (e.g., *reshenie*, *reshetka*, *napryazhenie*, in Russian, or *mode*, *stud*, *gutter*, in English) have no such information associated with them but are merely identified as microglossary entries, and, when they are looked up, they instruct the computer to check the location in the memory in which the microglossary code was stored at the beginning of the text. This code then directs the computer to the appropriate microglossary in which the word in question is looked up. The meaning

* This work has been partially sponsored under Contract AF30(602)-2617 by the Air Force Systems Command, Rome Air Development Center, Griffiss Air Force Base, New York.

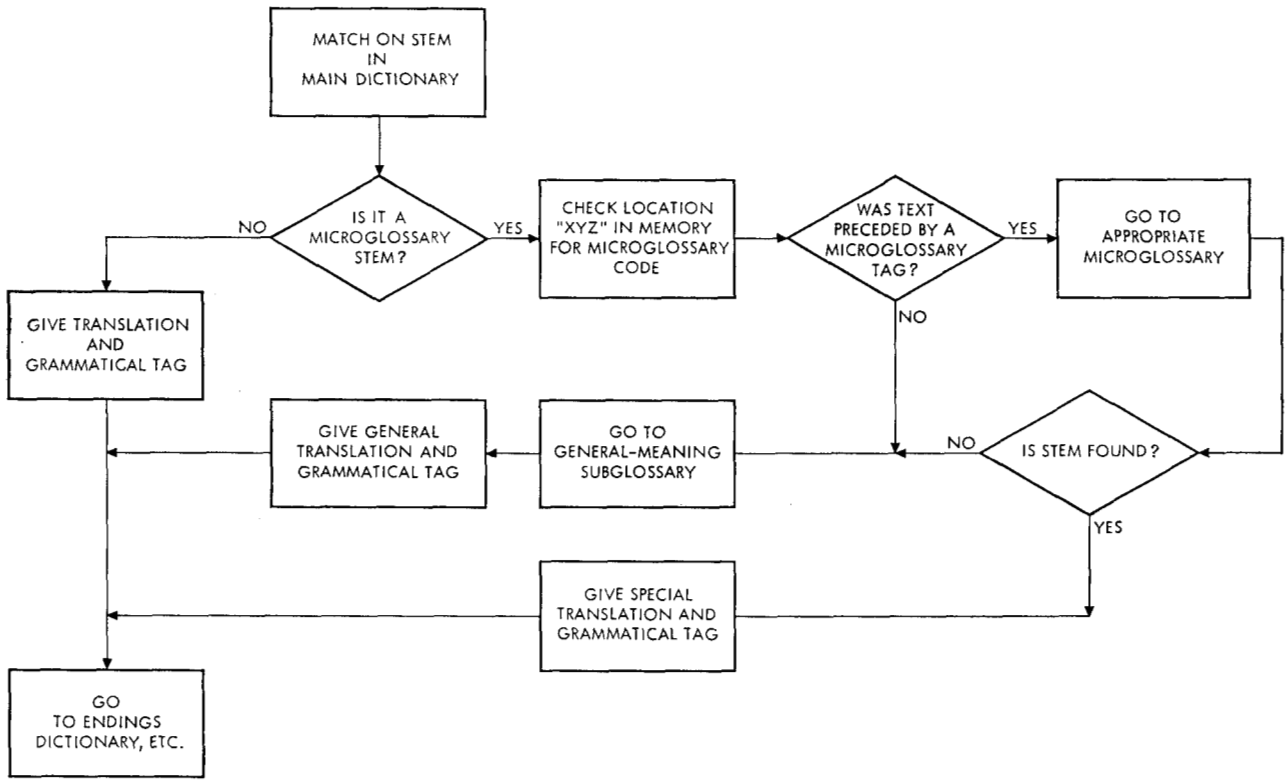

Figure 1 **Sequence of operations in a general-purpose automatic dictionary containing microglossaries.**

Table 1 **Sample of the IBM automatic Russian-English Dictionary containing microglossaries.**

Russian stem	*Instructions to the computer*	*Russian stem*	*Instructions to the computer*
Main dictionary		**Microglossaries**	
			MECHANICAL ENGINEERING
напряд	Translate as *spin*. Grammatical tag is G_1. Go to endings table.	напряжени	Translate as *voltage/stress*. Grammatical tag is G_7. Go to endings table.
напряжен	Translate as *tense*. Grammatical tag is G_2. Go to endings table.		
			ELECTRICAL ENGINEERING
напряжени	Microglossary entry. Check location *XYZ* in memory for *MG* code.	напряжени	Translate as *voltage*. Grammatical tag is G_7. Go to endings table.
напряженност	Translate as *intensity*. Grammatical tag is G_3. Go to endings table.		
.		.	
.		.	
		цепь	Translate as *circuit*. Grammatical tag is G_8. Go to endings table.
цепочк	Translate as *chain*. Grammatical tag is G_4. Go to endings table.		
			NUCLEAR PHYSICS
цеппелин	Translate as *zeppelin*. Grammatical tag is G_5. Go to endings table.	напряжени	Translate as *voltage/stress*. Grammatical tag is G_7. Go to endings table.
цепь	Microglossary entry. Check location *XYZ* in memory for *MG* code.		
		.	
церазин	Translate as *cerasin*. Grammatical tag is G_6. Go to endings table.	.	
		цепь	Translate as *circuit*. Grammatical tag is G_8. Go to endings table.
General meaning subglossary			
			PHYSICS
напряжени	Translate as *tension*. Grammatical tag is G_7. Go to endings table.	напряжени	Translate as *voltage/stress*. Grammatical tag is G_7. Go to endings table.
.		.	
.		.	
цепь	Translate as *chain*. Grammatical tag is G_8. Go to endings table.	цепь	Translate as *chain/circuit*. Grammatical tag is G_8. Go to endings table.

and other information given here are those which are the most probable in the given field. If the code corresponding to the "general dictionary" is found, i.e., if no microglossary tag occurred at the beginning of the text, the search is directed to a subglossary in which all the microglossary words are stored with the general translations.

Not every microglossary contains all the microglossary words, but only those words which have a special meaning in that field. If a word is not found in a microglossary for a given field, the search is directed to the subglossary which provides a general translation for the word in question. So, for instance, the Russian word *tsep*, which means *chain* or *circuit*, would be stored in the microglossaries for, say, electrical engineering and physics, where it would be given the translation *circuit*. It would also be stored in the general-meaning subglossary, but with the translation *chain*. No entries would be listed for this word in microglossaries for biology, mathematics, etc., since the translation *chain* would be provided for these fields by the general-meaning subglossary.

The fact that each word belonging to the microglossary category must be searched in either two or three dictionaries does not appreciably reduce the over-all translation speed, since the frequency of occurrence for words with specific meaning in various fields is, as a rule, fairly low. In this approach the most frequent words (i.e., words which are common to all the fields and are usually nontechnical) are not stored in microglossaries and are found during the first search in the general-purpose dictionary.

References

1. Victor A. Oswald, Jr., *Microsemantics*, mimeographed, June, 1952, as cited by Booth and Locke, Ref. 2.
2. A. D. Booth and W. N. Locke, "Historical Introduction" in *Machine Translation of Languages*, The Technology Press of MIT, John Wiley and Sons, Inc., N. Y., and Chapman and Hall, Ltd., London, 1955.
3. L. E. Dostert, "The Georgetown-IBM Experiment" in *Machine Translation of Languages*.
4. A. Jamotis and H. H. Josselson, "Multiple Meaning in Machine Translation", *Reprints of International Conference on Machine Translation of Languages and Applied Language Analysis*, Vol. II, Her Majesty's Stationery Office, London, 1962.

Received July 19, 1963