[*Eighth Annual Round Table Meeting on Linguistics and Language Studies*, Georgetown University, 1957]

Linguistics And Information Retrieval

Simon M. Newman
Patent Research Expert
U. S. Patent Office

Retrieval of information from technical literature as practiced in the Patent Office is much more concerned with the specific *interrelations* of *things* than with the specificity of the things themselves.

In this literature, any two documents which refer to the same or to closely similar phenomena will ordinarily express these phenomena in two widely different sets of verbal expressions. A patent examiner who may be searching for a similar phenomenon in formulating a retrieval request will probably utilize a still different, but accurate, verbal expression. Hence, in order to retrieve the documents, both the request and the documents must be transliterated into a common linguistic denominator.

We have been coining terms for an "unambiguous" or *Ruly English*, (as distinguished from "unruly",) after the terminology of Professor Dodd[1].

Our method of setting forth a disclosure, including the relationships found therein, has been termed *itemization.* In this system, each thing is listed as a separate item. Additional *descriptors* which state other aspects of each thing are listed in the same item. Relationship between items is shown by means of *interrelational concepts,* coined in mirror-image form, one of the images being *distributed* (listed) with each of the things being related. Actions are also distributed with each of the things involved. Distributed terms are identified by identical numbers, called *interfixes.*

We are coining Ruly English *roots* to be used as descriptors. These roots can be *modulated* to indicate such notions as process, thing processed, thing processing, condition and number. Ruly roots of action words are coined to encompass changes in all the things taking part in the action. Ruly roots of *qualifiers* likewise are coined to encompass *complementary* terms. A simple solution to the *quantification* of the qualifiers is proposed.

The Linguistic Approach

One of the approaches taken by our Office in its research on storage and retrieval of information in technical literature is linguistic in nature. This linguistic investigation was undertaken for two reasons. Most scientific and technical literature is already verbally expressed, and it is clear that other references in the form of drawings, tables of figures, photographs, models, working machinery, etc., could be transliterated into their verbal counterparts. It is also quite clear that no matter what final scheme or method of information retrieval is ultimately developed, an "unambiguous" metalanguage will be found to be of value.

Since conventional linguistic analysis did not prove to be helpful in the solution of our technical problems, we found it necessary to undertake steps to begin the creation of a metalanguage in which each unit will have one, and only one meaning, and in which each meaning may be expressed with one, and only one unit.[2] This metalanguage and the method proposed for its use undoubtedly will appear unconventional to trained linguists; however experimentation to date points to the eventual solution of some Patent Office problems in information retrieval.

We have named this metalanguage *Ruly English* after the terminology of Professor Dodd[1], who has pointed out that English is quite "unruly". Such a metalanguage will have at least two mutually exclusive uses in information retrieval.

Mechanized Information Retrieval

In our work at the Patent Office we are much more interested in the interrelation of *things* than we are in the specific details of the individual things themselves. Any two technical documents which refer to the same or to closely similar phenomena will ordinarily express these phenomena in two widely different sets of verbal expressions, both of which accurately convey the same information to the human mind. In using a mechanized searching system, a patent examiner might formulate his *request* for the retrieval of this same information by utilizing a still different, but accurate verbal expression. If his search is to yield both documents, the request and the documents must each be transliterated into a metalinguistic common denominator. This metalanguage for the first time will make possible the conversion of the many complex and interrelated notions in a particular document into single unique forms. Such a scheme as this is fundamental in the development of any successful retrieval system.

In an automatic assembly program system for accepting search request data, these same unique terms may be used to encode such data in a form suitable for machine instruction or command. The utilization of such systems is being widely developed by the machine industry at this moment.

Once the terminology and definition of the language units have been created, the encoding of search questions appears to be merely a problem of program development.

It also appears to be quite feasible to program a data processing machine to communicate with its operator during a search, in order to inform him of the trends of the results, so that he may vary the questioning program, or may substitute a different form of program or question before completing the search. This same metalanguage, thus, would serve as the means by which such intercommunication could be effected.

Basic Elements Of Ruly English

The basic elements of Ruly English are *itemization, distribution* and *intermixing. Itemization* consists of assembling all the descriptors of a single thing, each of these descriptors describing the thing from a different aspect, and grouping them as one item in a numbered list. Such numbers are used only for identification. *Distribution* is applicable to notions of interaction or interrelation between two or more items. Such notions are expressed by means of two or more *cognate descriptors,* one of which is placed with each related item. Thus distributed, the notion expresses the relation between the items. *Interfixing* is a method of tying together the cognate descriptors of distributed notions by affixing to each descriptor the same arbitrary number.

The Interrelational Concept

The notions of interrelations between items which are conveyed in English by prepositions were found to be quite troublesome. We therefore went to the list of *Basic English* words and took twenty-five which we recognized as prepositions. We collected all the phrases or sentences that we could find which utilized these prepositions, and we attempted to separate them into their various meanings. Fig. 1 is a list of the different meanings of *from.*

1. "differ" from - We distinguish day *from* night.
2. "means" from - The ball hangs *from* a string.
3. "cause" from - We get hives *from* berries.
4. "while" from - He throws *from a* standing position.
5. "start" from - She makes a cake *from* flour.
6. "reject" from - We shall appeal *from* a decision.
7. "exclude" from - We saved him *from* injury.
8. "away" from - They live far *from* a city.
9. "whence" from - The train came *from* New York; he took a penny *from* his pocket; the words are separated *from* context.

also

10. "whence" from ... to - He was measured *from* head *to* foot.

"FROM"S

Fig. 1

Next, we collated these phrases and sentences and collected those containing prepositions having equivalent meanings.

As shown in Fig. 2, we found that one meaning of each of *from*,

(1) As *from* in : The train came *from* New York; he took a penny *from* his pocket; the words are separated *from* context.

(2) As *on* in : He drew a check *on* the bank account; fighting the attack made an inroad *on* the supplies.

(3) As *off* in : He ate *off* the plate; he cut the end *off* the stick.

(4) As *of* in : He came *of* a noble family; wines *of* France are well known; art is bought *of* a dealer.

(5) As *from whence* in: Back to the dust *from whence* he came.

INTERRELATIONAL CONCEPT: WHENCEFROM(FROMWHENCE)

Fig. 2.

on, off, of, and the complex *from whence* all conveyed the same notion, which we arbitrarily named WHENCEFROM . Since this *interrelational concept* has a direction or polarity, in order to point out the direction of action, we created a *mirror image* of this concept, termed FROM-WHENCE. For example, the train (of example 1, Fig. 2) is FROM-WHENCE, and New York is WHENCEFROM. Not all interrelational concepts have polarity, e.g., the mirror image of AMONG is identical with its object term.

Itemization

Let us take the simple sentence, *The stale water is emptied from the china pitcher,* and note one way in which its meaning may be unambiguously presented. Fig. 3 shows the itemization of this sentence.

Item #	Unruly root	Ruly interrelational concept	Interfix
11	water		
	stale		
	empty	FROMWHENCE	15
12	pitcher		
	china		
	empty	WHENCEFROM	15

THE STALE WATER IS EMPTIED FROM THE CHINA PITCHER

Fig. 3

Item 11 is the water and the additional descriptor stale is included. Item 12 is the pitcher, and the additional descriptor *china* is included. The action *empty* has been distributed between both items, and the interrelated concept WHENCEFROM(FROMWHENCE) shows the direction of the emptying. The interfix 15 ties the two distributed cognate notions together.

This same idea could have been stated *The china pitcher is emptied of its stale water.* Using this same technique, we derive the itemization of Fig. 4 which is identical with Fig. 3, except for those factors which do not influence meaning, i.e., the order of the items, their numbers and the interfix number.

Item #	Unruly root	Ruly interrelational concept	Interfix
13	pitcher china		
	empty	WHENCEFROM	27
14	water stale		
	empty	FROMWHENCE	27

THE CHINA PITCHER IS EMPTIED OF ITS STALE WATER.

Fig. 4

Roots

We then decided that if we could reduce the elemental Ruly English terms to *Roots*[4],, we could modify the meaning by adding a *modulant.* Fig. 5 gives a partial list of modulants. For example, if the Root for

=NT	process
=W	work
=M	made from or out of
=E	condition
	Numerical
=X	or more
=Y	exactly
=Z	or less
=B	as an ordinal

MODULANTS

Fig. 5

the unruly notion *empty* is DISPEN (from the verb dispense) *the* action or process of emptying is DISPEN=NT, the condition of being emptied is DISPEN=E, etc. Numbers may also be modulated, e.g., 3=X for *three or more;* 3=Y for *exactly three;* and 3=B for *third.*

In choosing terms for Roots, many of the common names of things must be avoided. Such names often suggest a mere accidental use to which the thing is put, or some property incidental to its use. A *tray* may be inverted and become a *cover. A* plastic water *glass* or a piece of melamine dinner *china* are contradictions in themselves. From common usage we have chosen only such terms as *ring* or *sphere* which connote structural shape.

Modulants

As we create terms for our Ruly vocabulary, we must define them in part in unruly English, and, if possible, give examples to help interpret each term. For instance, CONFORM (from conformation) may be defined as *a geometric shape or figure in zero, one, two or three dimensions; e.g., a point, a line, a surface or a volume. As* more terms are defined, we can go back over our unruly definitions and substatute Ruly terms. Fig. 6 illustrates the Ruly definition and its unruly transliteration. The Root chosen is POLAR and it is defined in terms of its condition modulant, POLAR=E.

Ruly definition	Unruly transliteration
POLAR=E is	The condition of polarity is
given by a contrasting	given by a contrasting
characteristic of	characteristic of
1=Y ELEMENT or PORTION	one or more elements or portions
of a CONFORM	of a conformation
distinguishing it from	distinguishing it from
other ELEMENTS or PORTIONS	other elements or portions
of the same CONFORM	of the same conformation
e.g., the front of	e.g., the front of
a piece of furniture	a piece of furniture

POLAR (from polarity)

Fig. 6

Dual Aspect Roots

Action words often must be radically changed. For example, consider the situation in which one heats a room by cooling water - the conventional hot water heating system. This process is *dual-aspect*

in that the temperature of the air in the room is raised while that of the water is cooled. For this situation we coin the Root HEATCOOL, and as a *ground rule* we apply all interrelations to the first part of the compound Root, i.e., HEAT-.

This situation has been itemized in Fig. 7, in which we note that

Item #	Root and modulant	Interrelational concept and interfix
15	room HEATCOOL=NT	MORE-31
16	water HEAT=E HEATCOOL=NT	LESS-31

HEATING A ROOM BY COOLING HOT WATER

Fig. 7

the room increases in temperature because the process HEATCOOL= NT has the cognate image MORE of the interrelated concept MORE-(LESS). Conversely the HEAT=E (heated) water is cooled because the cognate image LESS applies to the HEATCOOL=NT.

Qualification of Quantifiers

Qualifying terms used in technical documents are ambiguous and troublesome. Both the balance spring of a watch and an enormous bridge girder may be designated as *flexible.* Actually analysis will show that all things are located somewhere on a scale between rigid and flexible, and none are found at either end of the scale. Hence we coin a *dual qualifier* such as RESILRIG to describe the condition, and we quantify it by one of two terms, *slightly* (SLI') or *substantially* (SUB'). Again we use the ground rule and apply the quantifier to the first part of the compound Root, RESIL-. Hence the bridge girder is SLI'RESILRIG and the balance spring is SUB'RESILRIG. If quantification can be measured along some scale, e.g., the Brinell standard of hardness, some predetermination will be made as to what portions are slightly hard or substantially hard, and it may be necessary to use both descriptors for values in the middle of the scale.

The one phase of research and development work expressed in this brief summary should provide insight into the linguistic problem which confronts the Patent Office. It follows that current research in the methodology of machine translation is, by extension, a contribution to the furtherance of our project. In proposing to adapt some of the techniques of machine translation to our specific uses, we become allies of your cause. Can we reciprocate in some measure with efforts that are useful in your research?

REFERENCES

1. Stuart C. Dodd, "Model English," in W. N. Locke and A. D. Booth, *Machine Translation of Languages*, pp. 167-73.
2. In such metalanguages as *Interglossa*, on the other hand, several units may convey the same meaning.
3. All *Ruly English* terms will be written in upper case letters.
4. Ruly English roots will be denoted Roots, i.e., with an upper case R.