DEARBORN OBSERVATORY CONTRIBUTIONS

NO. 37

AN AUTOMATIC QUESTION-ANSWERING SYSTEM FOR STELLAR ASTRONOMY

JACQUES F. VALLEE AND J. ALLEN HYNEK

Reprinted from the *Publications of the Astronomical Society of the Pacific*
(Vol. 78, No. 463, August 1966)

AN AUTOMATIC QUESTION-ANSWERING SYSTEM FOR STELLAR ASTRONOMY[*]

JACQUES F. VALLEE AND J. ALLEN HYNEK

Technological Institute and Lindheimer Astronomical Research Center, Northwestern University

Received February 14, 1966

Questions of a technical nature pertaining to the field of stellar evolution have been answered entirely by automatic means. In the research described, an English text submitted by the astronomer, consisting of a series of questions in machine-readable form, is scanned by a CDC 3400 data-processing machine and analyzed semantically. An information retrieval system automatically triggered by this analysis initiates a search through a star catalog and gives a numerical answer.

Introduction

In the last decades, the problem of storage and retrieval of large amounts of information has become a central one in many scientific disciplines, particularly those where heavy reliance is placed on statistical data. In stellar astronomy, for instance, the situation frequently arises in which catalogs need to be addressed at high speed. The Dearborn Observatory version (1965) of the Bright Star Catalogue, which contains space velocities both relative to the sun and to the local standard of rest, and which is available in machine-readable form, is a case in point. One might wish to use such a catalog to answer questions like the following:

How many binary systems have common proper motion components?

What is the proportion of spectroscopic binaries among all main-sequence stars which have only one visual component and whose total space velocity is less than 50 km/sec?

What is the percentage of giants later than F8 whose speed is below 70 km/sec?

What is the number of binaries in the Bright Star Catalogue whose primary is earlier than A3?

How many bright stars are double-line spectroscopic binaries?

Ordinarily, sorting schemes or separate programming would be necessary to extract this information (Weller 1964). In the research

[*] *Dearborn Observatory Contributions* No. 37.

to be discussed in this paper, we have designed a computer system that processes descriptive English questions such as the above and provides the corresponding answers at the average rate of eight questions per minute.

To achieve this result, we have first developed a compact code to express with only four characters the position of a star on the HR diagram, its spectral type, its multiplicity configuration, and its total space velocity. Next, we designed an information retrieval system capable of recognizing search strategies specified by "retrieval formulae," i.e. logical expressions which indicate without ambiguity the properties of the stars to be extracted from the Catalogue. Finally, a language processor was created which would accept natural language inputs and translate them into search strategies.

Algorithms for Converting the Characteristics of Stellar Systems into Codes

The *multiplicity code* of a stellar system is a standardized description of its configuration by an integer where each digit represents one component. In 1962, when the Dearborn version of the Bright Star Catalogue was converted to punched cards, a three-digit multiplicity code, where each system was considered as a potential visual triple, was defined (Table I).

Using this code, we can now express in logical terms the configuration of any system, as defined. In the following we shall call the *configuration index* of a stellar system a symbol expressing the

TABLE I

MULTIPLICITY CODE

	A	B	C
	First Component	Second Component	Third Component
0		None	None
1	Single star	Single star	Single star
2	Single-line sp. bin.	Single-line sp. bin.	Single-line sp. bin.
3	Double-line sp. bin.	Double-line sp. bin.	Double-line sp. bin.
4		Close companion	Close companion
5	Triple spectro. syst.	CPM, close pair	CPM, close pair
6	Composite spectrum	CPM, double-line sp. b.	CPM, double-line sp. b.
7		CPM, single-line sp. b.	CPM, single-line sp. b.
8	Astrometric binary	CPM, single star	CPM, single star
9	Single star, no RV	None	None

to be discussed in this paper, we have designed a computer system that processes descriptive English questions such as the above and provides the corresponding answers at the average rate of eight questions per minute.

To achieve this result, we have first developed a compact code to express with only four characters the position of a star on the HR diagram, its spectral type, its multiplicity configuration, and its total space velocity. Next, we designed an information retrieval system capable of recognizing search strategies specified by "retrieval formulae," i.e. logical expressions which indicate without ambiguity the properties of the stars to be extracted from the Catalogue. Finally, a language processor was created which would accept natural language inputs and translate them into search strategies.

Algorithms for Converting the Characteristics of Stellar Systems into Codes

The *multiplicity code* of a stellar system is a standardized description of its configuration by an integer where each digit represents one component. In 1962, when the Dearborn version of the Bright Star Catalogue was converted to punched cards, a three-digit multiplicity code, where each system was considered as a potential visual triple, was defined (Table I).

Using this code, we can now express in logical terms the configuration of any system, as defined. In the following we shall call the *configuration index* of a stellar system a symbol expressing the

TABLE I

MULTIPLICITY CODE

	A	B	C
	First Component	Second Component	Third Component
0		None	None
1	Single star	Single star	Single star
2	Single-line sp. bin.	Single-line sp. bin.	Single-line sp. bin.
3	Double-line sp. bin.	Double-line sp. bin.	Double-line sp. bin.
4		Close companion	Close companion
5	Triple spectro. syst.	CPM, close pair	CPM, close pair
6	Composite spectrum	CPM, double-line sp. b.	CPM, double-line sp. b.
7		CPM, single-line sp. b.	CPM, single-line sp. b.
8	Astrometric binary	CPM, single star	CPM, single star
9	Single star, no RV	None	None

AN AUTOMATIC QUESTION-ANSWERING SYSTEM FOR STELLAR ASTRONOMY*

JACQUES F. VALLEE AND J. ALLEN HYNEK
Technological Institute and Lindheimer Astronomical
Research Center, Northwestern University

Received February 14, 1966

Questions of a technical nature pertaining to the field of stellar evolution have been answered entirely by automatic means. In the research described, an English text submitted by the astronomer, consisting of a series of questions in machine-readable form, is scanned by a CDC 3400 data-processing machine and analyzed semantically. An information retrieval system automatically triggered by this analysis initiates a search through a star catalog and gives a numerical answer.

Introduction

In the last decades, the problem of storage and retrieval of large amounts of information has become a central one in many scientific disciplines, particularly those where heavy reliance is placed on statistical data. In stellar astronomy, for instance, the situation frequently arises in which catalogs need to be addressed at high speed. The Dearborn Observatory version (1965) of the Bright Star Catalogue, which contains space velocities both relative to the sun and to the local standard of rest, and which is available in machine-readable form, is a case in point. One might wish to use such a catalog to answer questions like the following:

How many binary systems have common proper motion components?

What is the proportion of spectroscopic binaries among all main-sequence stars which have only one visual component and whose total space velocity is less than 50 km/sec?

What is the percentage of giants later than F8 whose speed is below 70 km/sec?

What is the number of binaries in the Bright Star Catalogue whose primary is earlier than A3?

How many bright stars are double-line spectroscopic binaries?

Ordinarily, sorting schemes or separate programming would be necessary to extract this information (Weller 1964). In the research

* *Dearborn Observatory Contributions* No. 37.

gravitational relationships observed between its members, by reference to an enumeration of all possible configurations. If (ABC) is the multiplicity code of a star, we can partition the set of values assigned to A, B, and C by defining the logical statements and the correspondences between system configurations and logical representations listed in Table II.

It will be noted that configurations 10 and 11 are rare, peculiar systems. In the following we simply denote all triple systems by the symbol T. This leaves only ten non-triple configurations, which can be thus denoted by a single digit between 0 and 9.

Now, let c be the configuration index defined above. Let h and s be symbols of the HR region of the star and its spectral range, respectively, as shown in Table III. Let V be the total space velocity of the system with respect to the centroid of mass of nearby stars[*] and expressed in kilometers per second.

We define the "velocity index" of a star, and denote it by the symbol v, as follows:

$v = 0$ if V is unknown
$v = n$ if $10(n{-}1) < V \leq 10(n)$
$v = 9$ if $V > 80$

TABLE II

DEFINITION OF SYSTEM CONFIGURATIONS

Logical Propositions	c	System	Logical Representation	No. of Systems
P1: A=1 or 9 . . . "Primary is single"	0	0	P1 and Q1	6248
Q1: B=0 or 9 . . . "No secondary"	1	*	non-P1 and Q1	1172
Q2: B=1 or 4 . . . "Second. is single"	2	0-0	P1 and Q2	770
Q3: B=2 or 3 . . . "Sec. is spec. bin."	3	*-0	non-P1 and Q2	140
Q4: B=8 . . . "Sec. is CPM, single"	4	0-*	P1 and Q3	15
Q5: B=6 or 7 . . . "Sec. is CPM, sp. bin"	5	*-*	non-P1 and Q3	16
R1: C=0 or 9 . . . "System not triple"	6	0—0	P1 and Q4	334
Q6: B=5 . . . "Sec. is visual bin."	7	*—0	non-P1 and Q4	83
Symbols:	8	0—*	P1 and Q5	7
0 is a single star	9	*—*	non-P1 and Q5	4
* is a spectroscopic binary	10	0—0-0	P1 and Q6	2
0-0 is a visual pair	11	*—0-0	non-P1 and Q6	1
0—0 is a CPM pair	12	TRIPLE	non-R1	141

[*] Space velocity components were computed by the Cracovian method (Przybylski 1962) for 5175 stellar systems.

TABLE III

HR INDEX AND SPECTRAL RANGE

		MK	Spectral		Center
HR Region	h	Class	range	Spectral classes	
Unknown	0		$s = a$	earlier than B2	—B0
			b	B3 – B7	B5
Dwarf	1	V	c	B8 – A2	A0
			d	A3 – A7	A5
Giant	2	III	e	A8 – F2	F0
			f	F3 – F7	F5
Supergiant	3	I, II	g	F8 – G2	G0
			h	G3 – G7	G5
Subgiant	4	IV	i	G8 – K2	K0
			j	K3 – K7	K5
Hertzsprung Gap	5		k	K8 – M2	M0
			l	later than M2	M5—
Subdwarf	6	VI			

The word of information $\sigma(S) = hscv$ is called the "signature" of a stellar system S.

The "signature" is of value in providing a compact, mnemonic summary of the outstanding characteristics of a stellar system. For instance, we see that $1c02$ is the signature of a main-sequence star whose spectrum is in the range B8 to A2 and whose space velocity is between 10 and 20 km/sec. Similarly, $2k30$ refers to a two-component system where the brightest star is a late K or early M giant and a spectroscopic binary, while the secondary is a single star, the space motion of the system being unknown.

Retrieval Formulae

In the context of this paper, any question regarding the characteristics of a stellar population can be of one of two types: one either asks for a number, namely the cardinality of a certain subset of the Catalogue, or the percentage or fraction of objects having certain properties among a certain sub-population whose cardinality may or may not have been previously determined.

We recognize the importance of this difference between the two types of questions by calling them "questions of type n," "questions of type p," respectively.

Example of a question of type n:

"What is the number of stellar systems whose space velocity is inferior to 20 km/sec?"

TABLE III

HR INDEX AND SPECTRAL RANGE

		MK	Spectral		Center
HR Region	*h*	Class	range	Spectral classes	
Unknown	0		$s = a$	earlier than B2	—B0
			b	B3 – B7	B5
Dwarf	1	V	*c*	B8 – A2	A0
			d	A3 – A7	A5
Giant	2	III	*e*	A8 – F2	F0
			f	F3 – F7	F5
Supergiant	3	I, II	*g*	F8 – G2	G0
			h	G3 – G7	G5
Subgiant	4	IV	*i*	G8 – K2	K0
			j	K3 – K7	K5
Hertzsprung Gap	5		*k*	K8 – M2	M0
			l	later than M2	M5—
Subdwarf	6	VI			

The word of information $\sigma(S) = hscv$ is called the "signature" of a stellar system S.

The "signature" is of value in providing a compact, mnemonic summary of the outstanding characteristics of a stellar system. For instance, we see that 1*c*02 is the signature of a main-sequence star whose spectrum is in the range B8 to A2 and whose space velocity is between 10 and 20 km/sec. Similarly, 2*k*30 refers to a two-component system where the brightest star is a late K or early M giant and a spectroscopic binary, while the secondary is a single star, the space motion of the system being unknown.

Retrieval Formulae

In the context of this paper, any question regarding the characteristics of a stellar population can be of one of two types: one either asks for a number, namely the cardinality of a certain subset of the Catalogue, or the percentage or fraction of objects having certain properties among a certain sub-population whose cardinality may or may not have been previously determined.

We recognize the importance of this difference between the two types of questions by calling them "questions of type n," "questions of type p," respectively.

Example of a question of type n:

"What is the number of stellar systems whose space velocity is inferior to 20 km/sec?"

gravitational relationships observed between its members, by reference to an enumeration of all possible configurations. If (ABC) is the multiplicity code of a star, we can partition the set of values assigned to A, B, and C by defining the logical statements and the correspondences between system configurations and logical representations listed in Table II.

It will be noted that configurations 10 and 11 are rare, peculiar systems. In the following we simply denote all triple systems by the symbol T. This leaves only ten non-triple configurations, which can be thus denoted by a single digit between 0 and 9.

Now, let c be the configuration index defined above. Let h and s be symbols of the HR region of the star and its spectral range, respectively, as shown in Table III. Let V be the total space velocity of the system with respect to the centroid of mass of nearby stars[*] and expressed in kilometers per second.

We define the "velocity index" of a star, and denote it by the symbol v, as follows:

$v = 0$ if V is unknown
$v = n$ if $10(n-1) < V \leqq 10(n)$
$v = 9$ if $V > 80$

TABLE II

DEFINITION OF SYSTEM CONFIGURATIONS

Logical Propositions	*c*	System	Logical Representation	No. of Systems
P1: A=1 or 9 . . . "Primary is single"	0	0	P1 and Q1	6248
Q1: B=0 or 9 . . . "No secondary"	1	*	non-P1 and Q1	1172
Q2: B=1 or 4 . . . "Second. is single"	2	0-0	P1 and Q2	770
Q3: B=2 or 3 . . . "Sec. is spec. bin."	3	*-0	non-P1 and Q2	140
Q4: B=8 . . . "Sec. is CPM, single"	4	0-*	P1 and Q3	15
Q5: B=6 or 7 . . . "Sec. is CPM, sp. bin"	5	*-*	non-P1 and Q3	16
R1: C=0 or 9 . . . "System not triple"	6	0—0	P1 and Q4	334
Q6: B=5 . . . "Sec. is visual bin."	7	*—0	non-P1 and Q4	83
Symbols:	8	0—*	P1 and Q5	7
0 is a single star	9	*—*	non-P1 and Q5	4
* is a spectroscopic binary	10	0—0-0	P1 and Q6	2
0-0 is a visual pair	11	*—0-0	non-P1 and Q6	1
0—0 is a CPM pair	12	TRIPLE	non-R1	141

[*] Space velocity components were computed by the Cracovian method (Przybylski 1962) for 5175 stellar systems.

According to the velocity code we have defined (fourth digit of the "signature") we need the number of code-words which have as their last digit either a one (1) or a two (2). The answer is the sum of the cardinality of two subsets, and the retrieval formula will reflect this observation.

Example of a question of type p:

> "What is the proportion of spectroscopic binaries among all main-sequence systems which have only one component and whose total space velocity is less than 50 km/sec?"

This question calls for two main operations. If we had to find the answer by manual sorting of cards, we would:

a) Decide to neglect the second digit of the signature, since the question does not involve the spectral types.

b) Reject from the Catalogue all codes which do not begin with a one (1) since we are only concerned with the main sequence (see Table III).

c) Reject all codes which do not have either a zero (0) or a one (1) as configuration index, since we want systems with one visual component only (Table II).

d) Select all codes whose last digit is 1, 2, 3, 4, or 5 (see definition of the velocity index).

After this series of operations, we would be left with a subset of the Catalogue, containing signatures of the following types:

1.01	1.02	1.03	1.04	1.05
1.11	1.12	1.13	1.14	1.15

We have used a period (.) to indicate that the second digit (spectral range) was to be ignored. In order to answer the question we now have to evaluate the cardinality of the ten subsets defined by these formulae, then eliminate all codes that do not have a one (1) as third digit, keeping only the spectroscopic binaries. The final answer is the ratio of the number of systems in the latter set to the preceding number.

The procedure that we follow in establishing this sorting scheme can be clarified by the following notation: We shall denote by $Y = (abcd)$ the cardinality of the subset of the Catalogue defined by considering only those stellar systems whose signature is *abcd*. The two questions used as examples can now be restated in terms of this notation.

$Y = (...1) + (...2)$ answers the question "What is the number of stellar systems whose space velocity is inferior to 20 km/sec?" while

$$Y = (..1.)/((1.01) + (1.02) + (1.03) + (1.04) + (1.05) + (1.11) + (1.12) + (1.13) + (1.14) + (1.15))$$

defines the solution to the question we have just discussed in detail.

Such expressions are called *retrieval formulae.* It is natural to have a computer process these formulae and carry out the corresponding numerical manipulations, which are defined now without ambiguity. In this sense, retrieval formulae are a device for the mathematical representation of *search strategies.* An adequate formalism for the treatment of this problem is found in the theory of pushdown-store automata (to be published elsewhere by G. K. Krulee and Vallee).

The Linguistic Problem

In translating astronomical questions, expressed in ordinary English, into retrieval formulae, the difficulties we encounter are considerably smaller than those found in the general problem of machine translation. In our case, the object of language processing is to construct associations between descriptions of physical situations (expressed in the technical language of astronomy) and a machine configuration that extracts from the input statements formalized elements to be used in subsequent logical manipulations. The purpose of these manipulations is twofold:

1) To search for elements having certain properties, which should be extracted from the available records.
2) To perform certain operations on these subsets in order to answer a specific question.

Our solution to this problem is a program called ALTAIR (Automatic Logical Translation And Information Retrieval), where the input set of verbal statements has limitations which are ordinary in systems of this type. The operators *and, or, not* are prohibited, with the exception that the operator *and* may be used in ALTAIR within the list of attributes of a certain subset. Similarly, terms such as two questions used as examples can now be restated in terms of this notation.

$Y = (...1) + (...2)$ answers the question "What is the number of stellar systems whose space velocity is inferior to 20 km/sec?" while

$$Y = (..1.)/((1.01) + (1.02) + (1.03) + (1.04) + (1.05) + (1.11) + (1.12) + (1.13) + (1.14) + (1.15))$$

defines the solution to the question we have just discussed in detail.

Such expressions are called *retrieval formulae.* It is natural to have a computer process these formulae and carry out the corresponding numerical manipulations, which are defined now without ambiguity. In this sense, retrieval formulae are a device for the mathematical representation of *search strategies.* An adequate formalism for the treatment of this problem is found in the theory of pushdown-store automata (to be published elsewhere by G. K. Krulee and Vallee).

The Linguistic Problem

In translating astronomical questions, expressed in ordinary English, into retrieval formulae, the difficulties we encounter are considerably smaller than those found in the general problem of machine translation. In our case, the object of language processing is to construct associations between descriptions of physical situations (expressed in the technical language of astronomy) and a machine configuration that extracts from the input statements formalized elements to be used in subsequent logical manipulations. The purpose of these manipulations is twofold:

1) To search for elements having certain properties, which should be extracted from the available records.
2) To perform certain operations on these subsets in order to answer a specific question.

Our solution to this problem is a program called ALTAIR (Automatic Logical Translation And Information Retrieval), where the input set of verbal statements has limitations which are ordinary in systems of this type. The operators *and, or, not* are prohibited, with the exception that the operator *and* may be used in ALTAIR within the list of attributes of a certain subset. Similarly, terms such as

According to the velocity code we have defined (fourth digit of the "signature") we need the number of code-words which have as their last digit either a one (1) or a two (2). The answer is the sum of the cardinality of two subsets, and the retrieval formula will reflect this observation.

Example of a question of type p:

> "What is the proportion of spectroscopic binaries among all main-sequence systems which have only one component and whose total space velocity is less than 50 km/sec?"

This question calls for two main operations. If we had to find the answer by manual sorting of cards, we would:
a) Decide to neglect the second digit of the signature, since the question does not involve the spectral types.
b) Reject from the Catalogue all codes which do not begin with a one (1) since we are only concerned with the main sequence (see Table III).
c) Reject all codes which do not have either a zero (0) or a one (1) as configuration index, since we want systems with one visual component only (Table II).
d) Select all codes whose last digit is 1, 2, 3, 4, or 5 (see definition of the velocity index).

After this series of operations, we would be left with a subset of the Catalogue, containing signatures of the following types:

1.01   1.02   1.03   1.04   1.05
1.11   1.12   1.13   1.14   1.15

We have used a period (.) to indicate that the second digit (spectral range) was to be ignored. In order to answer the question we now have to evaluate the cardinality of the ten subsets defined by these formulae, then eliminate all codes that do not have a one (1) as third digit, keeping only the spectroscopic binaries. The final answer is the ratio of the number of systems in the latter set to the preceding number.

The procedure that we follow in establishing this sorting scheme can be clarified by the following notation: We shall denote by $Y = (abcd)$ the cardinality of the subset of the Catalogue defined by considering only those stellar systems whose signature is *abcd*. The

TABLE IV

SAMPLE OF QUESTION-AND-ANSWER PERIOD

Problems and solutions	Cumulative Execution Time (minutes)
Q. COMPUTE THE NUMBER OF SYSTEMS HAVING THE HR REGION UNKNOWN. Y=(0...).	0.00
A. THIS NUMBER IS 2298	
Q WHAT IS THE PROPORTION OF SYSTEMS HAVING STARS.	0.04
A. 8933 STARS OUT OF 8933, OR 100.0 PERCENT	
Q. REQUEST ALL STARS REDDER THAN F8 AMONG MAIN SEQUENCE SYSTEMS.	0.04
A. QUESTION NOT COMPATIBLE WITH CURRENT CODING SYSTEM AND/OR SEMANTICALLY INCORRECT	
Q. REQUEST ALL MAIN SEQUENCE SYSTEMS REDDER THAN F8. Y=((1G..)+(1H..)+(1I..)+(1J..)+(1K..)+(1L..))	0.04
A. THIS NUMBER IS 331	
Q. REQUEST ALL STARS BETWEEN F3 AND F7 HAVING SPEED ABOVE FORTY. Y=((.F.5)+(.F.6)+(.F.7)+(.F.8)+(.F.9))	0.25
A. THIS NUMBER IS 80	
Q. WHAT IS THE PROPORTION OF STARS HAVING A SPEED ABOVE 40 KM/SEC AMONG THE SYSTEMS WHOSE SPECTRAL TYPE IS BETWEEN F3 AND F7. Y=((...5)+(...6)+(...7)+(...8)+ (...9))/(.F..)	0.43
A. 80 STARS OUT OF 534, OR 15.0 PERCENT	
Q. COMPUTE THE NUMBER OF STARS OF MK CLASS III WHOSE SPECTRAL TYPE FALLS BETWEEN F3 AND K2. Y=((2F..)+(2G..)+(2H..)+(2I..))	0.49
A. THIS NUMBER IS 1463	
Q. WHAT IS THE PROPORTION OF SUBGIANTS AMONG THOSE STARS WHICH ARE OF SPECTRAL TYPE EARLIER THAN F7. Y=(4...)/((.A..)+(.B..)+(.C..)+(.D..)+(.E..)+(.F..))	0.63

```
A.  333 STARS OUT OF 4812, OR    6.9 PERCENT

Q. HOW MANY SYSTEMS OF MK CLASS IV ARE                    0.88
   INCLUDED IN THE BRIGHT STAR CATALOGUE.
A. QUESTION NOT COMPATIBLE WITH CURRENT CODING
   SYSTEM AND/OR SEMANTICALLY INCORRECT

Q. HOW MANY MULTIPLE STARS ARE DOUBLE.                    0.88
   Y=( (..2.)+(..3.)+(..4.)+(..5.)+(..6.)+(..7.)+(..8.)+(..9.)+(..T.) )
A. THIS NUMBER IS 1513

Q.  HOW MANY STELLAR SYSTEMS HAVE PRIMARIES.              1.22
   Y=(....)
A. THIS NUMBER IS 8933

Q. WHAT IS THE PROPORTION OF SYSTEMS HAVING               1.29
   A VELOCITY IN EXCESS OF 80 KM/SEC AMONG
   ALL STARS WHICH HAVE A COMMON PROPER
   MOTION COMPANION.
   Y=(...9)/( (..6.)+(..7.)+(..8.)+(..9.) )
A. 8 STARS OUT OF  428, OR   1.9 PERCENT

Q. FIND THE PERCENTAGE OF VISUAL PAIRS AMONG              1.44
   ALL STELLAR SYSTEMS WHERE THE PRIMARY IS
   REDDER THAN F3 AND WHOSE VELOCITY IS BELOW
   20 KM/SEC.
A. RETRIEVAL FORMULA LENGTH EXCEEDED
```

most or *highest* are prohibited. Such restrictions, as pointed out by the developers of BASEBALL are minor ones and can easily be removed at a later stage of refinement of the technique. In ALTAIR, however, the input sentence is not restricted to the form of a single-clause question since it was designed specifically to process questions of type p.

We give as illustration in Table IV a series of questions of the type considered in this article, as well as the answers given by ALTAIR. The language processor is capable of generating retrieval formulae from English statements at the average rate of 250 formulae per minute. The searching time through the Catalogue, however, is longer, and the resulting performance of the program is about eight answers per minute. We have purposely included some ques-

```
A.  333 STARS OUT OF 4812, OR    6.9 PERCENT

Q. HOW MANY SYSTEMS OF MK CLASS IV ARE                 0.88
   INCLUDED IN THE BRIGHT STAR CATALOGUE.
A. QUESTION NOT COMPATIBLE WITH CURRENT CODING
   SYSTEM AND/OR SEMANTICALLY INCORRECT

Q. HOW MANY MULTIPLE STARS ARE DOUBLE.                 0.88
   Y=( (..2.)+(..3.)+(..4.)+(..5.)+(..6.)+(..7.)+(..8.)+(..9.)+(..T.) )
A. THIS NUMBER IS 1513

Q.  HOW MANY STELLAR SYSTEMS HAVE PRIMARIES.           1.22
   Y=(....)
A. THIS NUMBER IS 8933

Q. WHAT IS THE PROPORTION OF SYSTEMS HAVING            1.29
   A VELOCITY IN EXCESS OF 80 KM/SEC AMONG
   ALL STARS WHICH HAVE A COMMON PROPER
   MOTION COMPANION.
   Y=(...9)/( (..6.)+(..7.)+(..8.)+(..9.) )
A. 8 STARS OUT OF  428, OR    1.9 PERCENT

Q. FIND THE PERCENTAGE OF VISUAL PAIRS AMONG           1.44
   ALL STELLAR SYSTEMS WHERE THE PRIMARY IS
   REDDER THAN F3 AND WHOSE VELOCITY IS BELOW
   20 KM/SEC.
A. RETRIEVAL FORMULA LENGTH EXCEEDED
```

most or *highest* are prohibited. Such restrictions, as pointed out by the developers of BASEBALL are minor ones and can easily be removed at a later stage of refinement of the technique. In ALTAIR, however, the input sentence is not restricted to the form of a single-clause question since it was designed specifically to process questions of type p.

We give as illustration in Table IV a series of questions of the type considered in this article, as well as the answers given by ALTAIR. The language processor is capable of generating retrieval formulae from English statements at the average rate of 250 formulae per minute. The searching time through the Catalogue, however, is longer, and the resulting performance of the program is about eight answers per minute. We have purposely included some ques-

TABLE IV

SAMPLE OF QUESTION-AND-ANSWER PERIOD

Problems and solutions	Cumulative Execution Time (minutes)
Q. COMPUTE THE NUMBER OF SYSTEMS HAVING THE HR REGION UNKNOWN. Y=(0...).	0.00
A. THIS NUMBER IS 2298	
Q WHAT IS THE PROPORTION OF SYSTEMS HAVING STARS.	0.04
A. 8933 STARS OUT OF 8933, OR 100.0 PERCENT	
Q. REQUEST ALL STARS REDDER THAN F8 AMONG MAIN SEQUENCE SYSTEMS.	0.04
A. QUESTION NOT COMPATIBLE WITH CURRENT CODING SYSTEM AND/OR SEMANTICALLY INCORRECT	
Q. REQUEST ALL MAIN SEQUENCE SYSTEMS REDDER THAN F8. Y=((1G..)+(1H..)+(1I..)+(1J..)+(1K..)+(1L..))	0.04
A. THIS NUMBER IS 331	
Q. REQUEST ALL STARS BETWEEN F3 AND F7 HAVING SPEED ABOVE FORTY. Y=((.F.5)+(.F.6)+(.F.7)+(.F.8)+(.F.9))	0.25
A. THIS NUMBER IS 80	
Q. WHAT IS THE PROPORTION OF STARS HAVING A SPEED ABOVE 40 KM/SEC AMONG THE SYSTEMS WHOSE SPECTRAL TYPE IS BETWEEN F3 AND F7. Y=((...5)+(...6)+(...7)+(...8)+ (...9))/(.F..)	0.43
A. 80 STARS OUT OF 534, OR 15.0 PERCENT	
Q. COMPUTE THE NUMBER OF STARS OF MK CLASS III WHOSE SPECTRAL TYPE FALLS BETWEEN F3 AND K2. Y=((2F..)+(2G..)+(2H..)+(2I..))	0.49
A. THIS NUMBER IS 1463	
Q. WHAT IS THE PROPORTION OF SUBGIANTS AMONG THOSE STARS WHICH ARE OF SPECTRAL TYPE EARLIER THAN F7. Y=(4...)/((.A..)+(.B..)+(.C..)+(.D..)+(.E..)+(.F..))	0.63

tions that were irrelevant to the problem context and were accordingly rejected by the program. Note that the questions were addressed to the Dearborn revised Bright Star Catalogue and were verified by reference to conventional sorting procedures.

As a linguistic processor, ALTAIR is still, of course, at an experimental stage. Implementation of the system requires access to a large computer, and our current tests are run on a CDC 3400. For most practical purposes, however, those wishing to take advantage of the flexibility of the scheme of *retrieval formulae* could very easily implement our method on practically any computer available on the market, by addressing directly the searching automaton. This would imply abandoning the idea of direct access in natural language and assumes that user has been trained to write retrieval formulae. The same scheme could readily be applied to any astronomical catalog by defining an appropriate system of codes similar to our system of "signatures."

Conclusion

We have tried to demonstrate the feasibility of an inquiry system driven by questions in ordinary English. It gives the astronomer direct access to Catalogue data and frees him from the necessity of wording his questions in the language of a particular data-processing system. Because of the large computer required, such a system will find its greatest efficiency in a time-sharing environment.

This research is supported as part of an interdisciplinary project conducted jointly by the astronomy department and the department of industrial engineering and management science at Northwestern University. The writers are indebted to Professor Gilbert K. Krulee for his encouragement and valuable guidance in designing this system.

REFERENCES

Dearborn Observatory Version of the Yale Bright Star Catalogue (1965) unpublished, available at the Observatory Library, Evanston.

Przybylski, A. 1962, *Acta Astronomica* 12, No. 4; *Mount Stromlo Obs. Reprint* No. 68.

Weller, W. J. 1964, *Pub. A.S.P.* 76, 152.