On the Application of Syntactic Methodologies in Automatic Text Analysis

Gerard Salton*
and
Maria Smith*

Department of Computer Science
Cornell University
Ithaca, NY 14853

Abstract

This study summarizes various linguistic approaches proposed for document analysis in information retrieval environments. Included are standard syntactic methods to generate complex content identifiers, and the use of semantic know-how obtained from machine-readable dictionaries and from specially constructed knowledge bases. A particular syntactic analysis methodology is also outlined and its usefulness for the automatic construction of book indexes is examined.

1. Introduction

It is generally agreed that new approaches must be introduced in information retrieval, if meaningful enhancements in retrieval effectiveness are to be obtained. New probabilistic and logical inference systems have in fact been proposed in the recent past that may provide useful, new models for retrieval activities, and suggestions have been made for using sophisticated methods of user-system interaction. Ultimately, any advanced information retrieval model must deal with the problem of *language analysis*, because the content of texts and documents necessarily controls the retrieval activities. This implies that methods must be available for analyzing the contents of documents and search requests, and for relating the information needs of users to the existing data bases. Such an analysis normally contains syntactic as well as semantic components.

Syntactic analysis systems have often been used to extract complex identifying units, such as noun phrases or prepositional phrases, from the texts of documents and search requests, and to distinguish useful content identifiers from more marginal ones. However, syntax by itself cannot resolve the many ambiguities that complicate the content analysis task. Additional contextual, or discourse-dependent, know-how is needed for that purpose. New aids are therefore normally considered, including especially term-defining information extracted from existing vocabulary schedules and dictionaries, and information obtained from manually prepared knowledge bases designed to reflect the semantic properties of particular areas of discourse.

The current situation in text analysis is briefly rviewed in this note, followed by an assessment of the limitations of purely syntactic methods for language analysis and index term generation.

2. Linguistic Approaches in Information Retrieval

A) Syntactic Analysis

Various attempts have been made in the recent past to use *syntactic analysis* methods for the identification of complex constructions--normally noun- and prepositional phrases--useful for the analysis of document and query text content. [1-5] Unfortunately, methods that are based on syntactic understanding alone are not sufficiently powerful to produce a proper analysis of available text samples:

- more often than not, several distinct analyses (parse trees) are obtained for particular text samples, and the resulting ambiguities are impossible to resolve by syntactic means;
- the vocabulary schedules used to provide information about the role and nature of individual words are often incomplete, and proper specifications necessary to carry out the syntactic analysis task may be lacking;

*This study was supported in part by the National Science Foundation under grant IRI 87-02735, and in part by a grant from OCLC Inc.

- the existing syntactic programs are often very large and demanding in terms of storage and computer power; for this reason, their use with large text samples is generally discouraged.

In practice, syntactic methods are thus often applied in a "fail-safe" manner, by pushing through the analysis of particular text samples even though full information may not be available for each text item, and certain grammatical rules may be violated by the available input. Shortcuts are then often used by restricting the syntactic methods to certain sample texts only--for example, the information queries received from the user population, but not the corresponding document texts, or the texts of only certain previously retrieved items. [4] Alternatively, a skimming-type parser may be used, that concentrates on certain text passages in preference to some others. [5]

Unfortunately, no matter how the problem is simplified, the analysis of noun phrase constructions, which is chiefly needed in information retrieval, is especially difficult, and all the various attempts to come up with general rules for noun phrase understanding have been unsuccessful. [6] Given a phrase such as "high frequency transistor oscillator", it is obvious that a simple rule, providing that the last word in the sequence (oscillator) is modified by all earlier components, fails in this case because the phrase does not refer to "high oscillators"; alternative rules suggesting that each term modifies the immediately adjacent one fail as well, because a "frequency transistor" is not a well-defined entity. Even if semantic information were available to eliminate the notion of a high (tall) transistor, or of a high (acute pitch) oscillator, the (false) notion of a high (tall) oscillator would still be difficult to reject. [6]

When syntactic methods are used for the generation of content-identifying phrases, the retrieval results are often discouragingly poor. Table 1 shows average search precision results obtained at ten recall points for two different collections of documents and queries. [7] In each case, the use of statistical (nonsyntactic) term generation methods is preferred over the syntactic analysis methods. Furthermore, for one of the two collections, the single-term indexing methods are preferred over the use of complex phrase identifiers.

B) Use of Dictionary Information

The failure of purely syntactic methods suggests that the generation of complex content identifiers may depend on the availability of additional information relating to the individual terms and to their interrelationships. One possibility consists in using the term descriptions contained in *machine-readable dictionaries* and thesauruses to improve the accuracy of term phrase formation. The thesaurus information may be used to disambiguate the meaning of terms and to generate groups of similar, or related, terms by identifying relationships among the contexts of various dictionary entries. [8-10]

Several attempts have been made to extract useful information from machine-readable dictionaries, and the experience indicates that some term relationships are relatively easy to obtain: notably certain synonym relations that are often explicitly identified in the dictionary, and hierarchical, taxonomic relations between terms that are identifiable following analysis of the dictionary definitions. [8] On the other hand, many complications also arise:

- many terms carry several defining statements in the dictionary, and the definition actually applicable in a given case may not be easily found;
- the printed definition may be difficult to parse, in which case the meaning of the defining statement may remain obscure;
- the relationships between different defining statements may be hard to assess.

Consider, as an example, the definitions for certain cryptographic terms extracted from Longman Dictionary of Contemporary English (1978 edition) shown in Table 2. It may be noted that the notion of "secret writing" or "secret message" occurs in definition 4 of cipher, definitions 1 and 2 of code, and finally definition 3 of key. An automatic identification of such overlapping definitions is, however, very difficult, and the subsequent formation of term cliques must remain elusive. The same is true for the analysis of many standard dictionary definitions. Given, for example, the notion that a robin is "a common type of fat little European bird with a brown back and wings and a red breast" (see Longman), an automatic parser must consider the possibility that a robin is a kind of fat, or a kind of fat European. Furthermore the modifier "brown" may apply only to "back", or additionally also to "wings" and "breast". Finally the preposition "with" may be used in its attributive sense, or possibly in the sense of containment (a cup with coffee and milk), or in various other senses.

Table 3 shows an evaluation of parsings of dictionary definitions obtained by Fox and coworkers. [8] The parsing accuracy varies between 60 and 77 percent, and in general several acceptable analyses are generated for each dictionary definition. These results indicate that dictionary information will not soon become generally usable in general text analysis systems.

C) Knowledge Base Construction

A great many attempts have been made to incorporate manually constructed knowledge bases for particular subject areas in automatic text analysis. [11-17] A knowledge base is an abstract representation of a topic area, or of a particular environment, including the main concepts of interest in that area, and various relationships between the entities. Several models are used for the representation of knowledge, including semantic networks, as well as collections of frames and scripts. When a knowledge base is available representing a particular subject area, the following extended retrieval strategies can be used:

i) the available search request is analyzed into a formal representation similar to that used for the knowledge base;

ii) a fuzzy matching operation is performed to compare the formalized search requests with elements of the knowledge representation;

iii) an answer to the search request is constructed if the degree of match between knowledge base and search requests is sufficiently great.

A typical example of such a matching operation is shown in Table 4 for a database of "Stardate" texts (giving information about stars and planets). [11] A typical frame specifying the relative position of two outer-space objects, designated as the "actor" and the "object" respectively, is identified as the "astro-pos-r1" frame in Table 4. A query, or input statement, matches this frame if the required frame components are also present in the input sentence, and if the relationships specified between frame components are maintained in the input. In the example of Table 4, the input "this past Christmas night, the moon (actor) was near (distance specification) Venus (object)" matches the corresponding frame.

As the example suggests, the use of knowledge bases in text processing environments raises many questions. There is first the problem of knowledge representation itself. It is normally impossible to isolate particular slices of knowledge in a self-contained way. That is, the interpretation of particular pieces requires not only the local subject knowledge, but also a wider context that is generally absent from particular knowledge bases. In that case, the text input cannot be properly compared with an incomplete knowledge framework. The example of Table 5 illustrates this fact for a knowledge base dealing with recorded acts of terrorism. [13] The context of various acts of terrorism that occurred in October 1988 shows that an interpretation of terrorist acts in Israel requires knowledge of Islamic fundamentalism, the Israeli elections, and the demographic situation in the Middle East. Similarly, terrorist activities in Algeria also involve the oil price situation within OPEC, and the rivalry between Algerian politicians. In general it is inconceivable that the varied contexts explaining particular facts or events should all be completely representable in existing knowledge bases. This implies that the matching possibilities between arbitrary inputs and existing knowledge bases will always be limited.

The other main problem with knowledge representations is the difficulty of analyzing and disambiguating the input texts and of transforming them into a form comparable to the available knowledge bases. This problem is illustrated in part by an example cited by Lewis et al. [17], where a searcher is interested in "research on failures in memory chips carried by high energy cosmic rays; also in microprocessors containing multipliers". A disambiguation of this statement involves analysis of the phrase "high energy cosmic rays" which is multiply ambiguous. Furthermore, it is unclear whether the user is interested in microprocessors alone, or in failures in microprocessors, and a decision about this question cannot be made without additional contextual information. Even when an input statement is correctly analyzed, the input vocabulary may be very different from the knowledge base specifications, and complicated inference rules may then be needed to relate such different forms of knowledge.

For the moment, only hand-constructed examples of knowledge-base utilization are available, and the knowledge-based approach has not proved itself in unrestricted text environments. However, simpler, language-based text analysis methods, not primarily founded on semantic considerations may in fact be useful in text processing environments. A syntactic text analysis method is examined in the remainder of this note, and its usefulness for indexing purposes is assessed.

3. Use of Syntax for Index Term Generation

A) Syntactic Analysis

It was noted earlier that sophisticated syntactic analysis systems now exist, that are capable of furnishing structural analyses for arbitrary samples of English input. Many of these analyzers operate with large grammars of several hundred grammatical rules, and provide output also for language fragments, such as book titles or section headings, that are not available in full sentence form. At the same time, it is clear that syntactic systems not based on contextual and other semantic considerations cannot be completely successful in eliminating ambiguities and constructing useful indexing units.

For the most part, syntactic analysis systems have been used with small text samples, or with special-purpose texts such as user queries. The effectiveness of syntactic systems is however best assessed by using large samples of general-purpose texts. The evaluation which follows is based on the analysis of a complete book chapter, consisting of over 7,000 words, dealing with the general topic of text compression. [18] The analysis system in use is the PLNLP English grammar (PEG) developed at the IBM Research Laboratory in Yorktown Heights for incorporation into the EPISTLE (or CRITIQUE) text critiquing system*. [19,20] This system analyzes complete sentences, as well as sentence fragments, producing in each case one or more syntactic parses for each sentence, ranked in decreasing order of presumed correctness. When the input cannot be analyzed using the normal grammar rules, a "fitted" parsing system is used in the PEG system to produce a reasonable analysis for the apparently intractable fragment. [20] In the experiments described in this section, only the top (most likely) parse tree for each sentence is used for index term generation.

The output of the syntactic analysis is used as input by a term phrase generation system designed at Cornell which generates single terms as well as term phrases suitable for indexing purposes. [7,21] The phrase generation system contains the following components:

- an element that recognizes sentence excerpts consisting of adjective and/or noun premodifiers, followed by a head noun, followed in turn by one or more postmodifiers.
- a system that identifies prepositional phrases using the preposition "of", and builds an inverted indexing entry in certain circumstances (for example, the phrase "retrieval of information" may be transformed into "information retrieval").
- a component that takes conjunctive constructions using the conjunction "and", or comma "," and distributes premodifiers and/or postmodifiers across the conjunction (thus "automatic indexing, enciphering, and decryption" produces the noun phrases "automatic indexing", "automatic enciphering", and "automatic decryption").
- a system of exclusion rules that prevents the phrase construction when the head noun appears on a list of prohibited words, and eliminates pre- and postmodifiers appearing on common word lists.
- a component that provides special treatment for certain constructions, including capitalized words, and words occurring in titles and section headings.

Basic statistical data for the syntactic analysis and phrase generation processes are included in Table 6. The Table shows that only about one third of the 318 sentences contained in the sample book chapter are perfectly analyzed by the PLNLP grammar. An example of a correctly analyzed sentence is reproduced in Fig. 1. The Figures shows a standard parse tree placed on its side, that is, rotated by ninety degrees in a counterclockwise direction. The first column of Fig. 1 represents the top node of the tree, showing that a declarative sentence (DECL) is being analyzed. The next column breaks down the sentence into prepositional phrase (PP), noun phrase (NP), verb (VERB), and adjective phrase (AJP). The prepositional phrase is further broken down in column 3 into a preposition (PREP), verb (VERB), and noun phrase (NP), and so on. The phrases generated for the sample input sentence

"by eliminating redundancies--a method known as text compression, it is often possible to reduce test sizes considerably without any loss of text content."

are shown at the bottom of Fig. 1. It may be noted that "text content loss" is produced by inversion from "loss of text content".

Beside the perfectly analyzed first third of the sentences, minor problems occur in the analysis of another third of the sentences. Minor errors may consist of erroneous classifications of particular words--for example, a noun analyzed as an adjective, or vice-versa--that may not seriously affect the phrase generation process. Adding the perfectly analyzed third of the sentences to those exhibiting minor problems, one finds that acceptable output is produced for about two-thirds of the input sentences. This confirms the success rate of dictionary parses reported earlier in Table 3.

Unfortunately, one-third of the input analyses are seriously flawed. An example of such an erroneous analysis is shown in Fig. 2 for the sentence

"today large disk arrays are usually available, but using short texts and small dictionary sizes saves processing time in addition to storage space and still remains attractive."

As the Figure suggests, there are multiple problems here, ranging from the misclassification of "today" as a noun, to the interpretation of "sizes" as the main verb (from "to size"), and of "saves" as a noun. The erroneous analysis generates a large number of false phrases, shown at the bottom of Fig. 2. Indeed, of the phrases shown in Fig. 2, only "storage space" is acceptable.

[*] The writer is grateful to the IBM Corporation and to Dr. George E. Heidorn for making available the PLNLP syntactic analysis system for use at Cornell University.

A summary of the phrase generation statistics is shown at the bottom of Table 6. The Table shows that about 85 percent of the phrases obtained by the phrase generation process may be acceptable. Fifteen percent of the phrases produced are clearly in error; in addition, the false analyses prevent the generation of many correct phrases, amounting to about 20 percent of the total number produced. In the example of Fig. 2, the correct phrases "short texts" and "dictionary sizes" cannot be obtained.

Table 7 contains a classification of phrase errors based on a total of 226 errors. By far the largest number of errors (over one-half) fall into the category of false syntactic word classifications. For example, words ending in "ing" function variously as nouns, adjectives, or verbs (present participles); correct classifications are very difficult to obtain in such cases. Another 35 percent of errors are due to misapplication of the phrase generation rules, such as the inversion rule for prepositional phrases with "of", the deletion of common words (that are not always common in every environment in which they occur), and the misapplication of the distribution of modifiers across conjunctions.

The illustrations provided in Table 7 and Figs. 1 and 2 confirm the trade-off for each phrase generation rule between cases where a rule proves beneficial, and other cases where the same rule does not work. On balance, the current rules provide useful phrase output. However, without deeper language understanding, it is impossible to prevent the occasional misfiring of any of the rules in certain contexts.

B) Book Indexing Application

The phrases generated by syntactic analysis can be collected to provide indexes for complete book chapters, or for sections within chapters. An example of a collected phrase list is shown in Fig. 3 for section 1 of the previously analyzed book chapter. [18] It is obvious that such a list is not immediately usable, not only because of the false phrases that are necessarily included (such as "communications psycholinguistics"), but also because of multiple occurrences of partly overlapping phrases, and a general lack of language normalization.

Various language normalization rules may be applied to the phrase output produced by the phrase generation rules. Such a normalization is designed to reject unusual term combinations that are often erroneous, while emphasizing term combinations that occur in many contexts, or that are produced by a variety of different rules. A typical set of phrase normalization rules usable for book index production is shown in Table 8. Special rules are provided in the Table for title phrases and for phrases with capitalized components; in addition, phrase matching methods are used to identify word combinations with multiple occurrences. By applying frequency thresholds that vary with text length, a final phrase index is automatically produced.

A typical automatic index obtained for the book chapter under analysis is shown on the left-hand side of Table 9. A corresponding manually built index is included for the same text material on the right side of Table 9. The output of Table 9 was obtained by taking title phrase, plus phrases that occurred sufficiently often, or had a sufficiently high partial-matching coefficient (see rule 2 of Table 8) in the sections of the chapter to warrant inclusion in the phrase index.

It may be noted that the automatic index of Table 9 contains only reasonable phrases, the obviously false term combinations having been eliminated by the normalization rules. Some phrases are highly germane (e.g. Huffman code, compression ratio, Zipf, etc.). Others are less compelling, including, for example, alphabetic characters, frequent characters, and text words. A comparison with the manual index reveals a large amount of overlap. However, certain useful phrases, such as coding efficiency, or language redundancy, are not included in the automatic product.

An evaluation of the foregoing phrase production system and language normalization procedures requires actual use in retrieval environments. The available experience indicates that the syntactic methodologies, supplemented by appropriate statistical and other normalization techniques, show much more promise in user environments where unrestricted natural-language texts must be processed than semantics-based artificial intelligence approaches. A precise determination remains to be made of the role and the effectiveness of the syntactic methodologies in automatic text manipulation systems.

References

[1] C. Berrut and P. Palmer, *Solving Grammatical Ambiguities within a Surface Syntactical Parser for Automatic Indexing*, Proc. of the Ninth International Conference on Research and Development in Information Retrieval, F. Rabitti, editor, Pisa, Italy, September 1986, 123-130.

[2] G. Thurmair, *A Common Architecture for Different Text Processing Techniques in an Information Retrieval Environment*, Proc. of the Ninth International Conference on Research and Development in Information Retrieval, F. Rabitti, editor, Pisa, Italy, September 1986, 138-143.

[3] A.F. Smeaton and C.J. van Rijsbergen, *Experiments on Incorporating Syntactic Processing of*

User Queries into a Document Retrieval Strategy, Proc. of the Eleventh International Conference on Research and Development in Information Retrieval, Y. Chiaramella, editor, Grenoble, France, June 1988, 31-51.

[4] A.F. Smeaton, *Incorporating Syntactic Information into a Document Retrieval Strategy: An Investigation*, Proc. of the Ninth International Conference on Research and Development in Information Retrieval, F. Rabitti, editor, Pisa, Italy, September 1986, 103-113.

[5] P.S. Jacobs and L.F. Rau, *Natural Language Techniques for Intelligent Information Retrieval*, Proc. of the Eleventh International Conference on Research and Development in Information Retrieval, Y. Chiaramella, editor, Grenoble, France, June 1988, 85-99.

[6] K. Sparck Jones and J.I. Tait, Automatic Search Term Variant Generation, *Journal of Documentation*, 40:1, March 1984, 50-66.

[7] J. Fagan, *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Nonsyntactic Methods*, Doctoral Dissertation, Cornell University, Report TR 87-868, Department of Computer Science, Ithaca, NY, September 1987.

[8] E.A. Fox, J.T. Nutter, T.Ahlswede, M. Evens and J. Markowitz, *Building a Large Thesaurus for Information Retrieval*, Proc. Second Conference on Applied Natural Language Processing, Association for Computational Linguistics, Austin, TX, February 1988, 101-108.

[9] M. Isoda, H. Aiso, N. Kamibayashi and Y. Matsunaga, *Model for Lexical Knowledge Base*, Proc. Eleventh International Conference on Computational Linguistics-Coling 86, University of Bonn, August 1986, 451-453.

[10] Y. Chiaramella, B. Defude, M.F. Bruandet and D. Kerkouba, *Iota: A Full Text Information Retrieval System*, Proc. of the Ninth International Conference on Research and Development in Information Retrieval, F. Rabitti, editor, Pisa, Italy, September 1986, 207-213.

[11] M. Mauldin, J. Carbonell and R. Thomason, *Beyond the Keyword Barrier: Knowledge-Based Information Retrieval*, Proc. 29th Annual Conference of National Federation of Abstracting and Information Services, Elsevier Press, 1987.

[12] U. Hahn and U. Reimer, *Informationslinguistische Konzepte der Volltextverarbeitung in TOPIC (Linguistic information concepts in the full text information processing system TOPIC)*, Report TOPIC 2/82, University of Konstanz, Germany, November 1982.

[13] R.M. Tong, L.A. Appelbaum, U.N. Askman and J.F. Cunningham, *Conceptual Information Retrieval Using RUBRIC*, Proc. of the Tenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, C.T. Yu and C.J. van Rijsbergen, editors, New Orleans, LA, June 1987, 247-253.

[14] D. DeJaco and G. Garbolino, *An Information Retrieval System Based on Artificial Intelligence Techniques*, Proc. of the Ninth International Conference on Research and Development in Information Retrieval, F. Rabitti, editor, Pisa, Italy, September 1986, 214-220.

[15] M.F. Bruandet, *Outline of a Knowledge Base Model for an Intelligent Information Retrieval System*, Proc. of the Tenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, C.T. Yu and C.J. van Rijsbergen, editors, New Orleans, LA, June 1987, 33-43.

[16] W.B. Croft and D.D. Lewis, *An Approach to Natural Language Processing for Document Retrieval*, Proc. of the Tenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, C.T. Yu and C.J. van Rijsbergen, editors, New Orleans, LA, June 1987, 26-32.

[17] D.D. Lewis, W.B. Croft and N. Bandharu, *Language Oriented Information Retrieval*, to be published in International Journal of Intelligent Systems, also COINS Technical Report 88-36, Department of Computer Science, University of Massachusetts, Amherst, MA, April 1988.

[18] G. Salton, *Automatic Tex Processing*, Addison-Wesley Publishing Co., Reading, MA, 1989, Chapter 5.

[19] G.E. Heidorn, K. Jensen, L.A. Miller, F.J. Byrd and M.S. Chodorow, The EPISTLE Text Critiquing System, *IBM Systems Journal*, 21:3, 1982, 305-326.

[20] K. Jensen, G.E. Heidorn, L.A. Miller and Y. Ravin, Parse Fitting and Prose Fitting: Getting Hold of Ill Formedness, *American Journal of Computational Linguistics*, 9:3-4, July-December 1983, 147-160.

[21] J. Kaye, Back-of-the Book Indexing Project, Term Report CS490, Department of Computer Science, Cornell University, Ithaca, NY, Spring 1988.

	CACM Collection 3204 documents 52 queries	CISI Collection 1460 documents 76 queries
Single-term indexing	0.2604	0.2450
Single terms and co-occurrence phrases	0.3195 (+22.7%)	0.2503 (+2.2%)
Single terms and syntactic phrases	0.2830 (+8.7%)	0.2480 (+1.2%)

Retrieval Performance of Syntactically Generated Phrases (from [7])
(average precision at 10 recall points)

Table 1

cryptogram (n)	1.	a *message* or *writing* in *secret* letter.
cipher (n)	1.	the number zero
	2.	any of the numbers from 0 to 9
	3.	a person of little importance and no influence
	4.	a system of *secret writing* (cf. code)
	5.	(rare) a personal sign made up of the first letters of one's names ...
code (n)	1.	a system of *secret* words, letters, numbers, etc. used ... to keep the *message secret*
	2.	a system of signals used instead of letters and numbers in a *message* that is to be broadcast ...
	3.	a body of established social customs
	4.	a collection of laws
key (n)	1.	an instrument, usually made of metal, for locking or unlocking ...
	2.	any one of the parts in a musical instrument, or in a machine, that is pressed down to produce the desired sound or other result
	3.	something that explains, answers, or helps you to understand (a key to the grammar exercises; to the *secret writing*)
	4.	someone or something that is very important
	5.	a set of musical notes with a certain starting or base note
	6.	a seed of certain types of tree ... which has winglike parts
	7.	"all in the same key" without any change of expression.

Typical Dictionary Definitions for Cryptographic Expressions
(from Longman Dictionary of Contemporary English)

Table 2

	Percent of Correctly Parsed Definitions	Number of Parses per Correctly Parsed Definition
2949 nouns	77.63%	1.70
1451 adjectives	68.15%	1.85
1272 information verbs	64.62%	1.59
2549 transitive verbs	60.29%	1.50

Performance of Syntactic Analysis of Dictionary Definitions
(from [8])

Table 3

Input Sentence:

"This past Christmas night,the moon was near Venus."

Formal Representation:

```
(is ("distance" (val(-8))))
(actor ("earths-moon"))
(object ("venus"))
(time ("time"))
```

Matching Frame in Knowledge Base:

```
astro-pos-r1

   (actor (required view-object "outer-space-object"))
   (is ("distance" (val (optional distance "number"))))
   (object (required position-center "outer-space-object"))
   (time (optional position-date "time"))
   (mode (optional position-mode "mode"))

(causals (infer astro-view-r1))

(constraints ((different (actor)(object)))
```

Comparison of Input Sentence with Knowledge Base Fragment
(example suggested by Michael Mauldin [11])

Table 4

Main Topic		Extraneous Contexts in Newspaper Stories
Terrorism in Israel	1.	islamic fundamentalism, conditions in Iran
	2.	demographic problems in Israel, fertility rates of Israeli and Arab populations
	3.	elections in Israel in November 1988 and different approaches of the Labor and Likud parties
Uprising in Algeria	1.	collapse of oil prices and general weakness of OPEC cartel
	2.	economic crisis in Algeria
	3.	islamic fundamentalism and its influence on Algerian population
	4.	rivalry between Algerian President Chadli and former President Ben Bella

Typical Context for Stories on Terrorism and Uprising

Table 5

Sample text	Chapter 5 on text compression of Automatic Text Processing [18]	
Total sentences	318 sentences consisting of 82 paragraphs or section headings	
Number of sentences	parsed with normal grammar	208
Number of sentences	parsed with fitted grammar	110
Sentence Statistics (total sentences 218)		
perfect syntactic analysis		103 (32%)
minor problems (not serious for indexing)		110 (35%)
major problems (will affect indexing)		105 (33%)
Phrase Statistics (total phrases 633)		
correct phrases		381 (60%)
marginal phrases		159 (25%)
false phrases		93 (15%)
phrases missed		133 (21%)

Phrase Formation Statistics

Table 6

Types of Errors	Examples
Syntactic classification (51%)	"converting (adj.) false text" "partial message consisting (noun)"
Inversion of prepositional phrase (12%)	"word occurrence half" (from "half of word occurrences")
Terms used as example (10%)	"word American" (from "the word American has 8 letters")
Wrong conjunctive analysis (9%)	"communications psycholinguistics" (from "communications theory and psycholinguistics")
Wrong term deletion (14%)	"word" (from "number of words")
False clause analysis (2%)	"reason substantial" (from "for this reason substantial (noun)//efforts(noun) were made")
Idioms (2%)	"addition" (from "in addition")

Errors in Phrase Formation

Table 7

Index Generation Rules

1. Occurrence frequency of phrases
   Use phrases (at least two components) exhibiting an occurrence frequency of at least *n*. Delete shorter phrases included in longer ones.

2. Partially matching phrases
   Use matching components of partly overlapping phrases to construct new phrases, provided that overlap consists of at least *n* matching phrase components. Remove shorter phrases included in longer phrases [*frequency* considerations + *frequency characteristics* + statistical language *characteristics* = frequency characteristics (overlap of 2 components)]

3. Capitalized phrases
   Use phrases with capitalized components with occurrence frequency of at least *p*. Remove longer phrases that cover shorter ones

4. Title phrases
   Keep phrases occurring in titles and section headings regardless of occurrence frequency.

5. Italicized phrases
   Keep phrases with italicized components, assuming that italics are automatically detectable.

Simplified Index Generation Rules

Table 8

	Automatic Index			Typical Manual Index
	alphabetic characters	(frequency 6, 3 sections)		coding efficiency
*	compression ratio	(frequency 10, 3 sections)	*	compression ratio
	English text	(frequency 2, capitalized)		differential coding
	English words	(frequency 2, capitalized)	*	entropy measurements
*	entropy measurements	(frequency 1, title)	*	fixed-length codes
*	fixed-length codes	(frequency 4, title)	*	Huffman code
	frequency considerations	(frequency 1, title)		information theory
	frequent characters	(frequency 6, 2 sections)		information value
*	Huffman code	(frequency 3, capitalized)		language redundancy
	occurrence probability	(frequency 7, 2 sections)		multicase coding
	special-purpose compression systems	(frequency 1, title)		numeric coding
				rank-frequency law
*	statistical language characteristics	(frequency 1, title)	*	statistical language characteristics
*	text compression systems	(frequency 1, title)	*	text compression
	text words	(frequency 6, 2 sections)	*	variable-length codes
*	variable-length codes	(frequency 1, title)		word frequency
*	word-fragment encoding	(frequency 1 title)	*	word-fragment coding
*	Zipf	(frequency 2, capitalized)	*	Zipf

Typical Automatically Produced Book Index (Chapter 5)
(* common entries for manual and automatic index)

Table 9

```
DECL    PP      PREP    "by"
                VERB*   "eliminating"
                NP      NOUN*     "redundancies"
                        PUNC      "--"
                        NAPPOS    DET       ADJ*      "a"
                                  NOUN*     "method"
                                  PTPRTCL   VERB*     "known"
                                            PP        PREP      "as"
                                                      NP        NOUN*       "text"
                                                      NOUN*     "compression"
                                                      PUNC      ","

        NP      PRON*   "it"
        VERB*   "is"
        AJP     AVP     ADV*      "often"
                ADJ*    "possible"
                INFCL   INFTO     "to"
                        VERB*     "reduce"
                        NP        NP        NOUN*"    "text"
                                  NOUN*     "sizes"
                                  PP        AVP       ADV*      "considerably"
                                            PREP      "without"
                                            QUANT     ADJ*      "any"
                                            NOUN*     "loss"
                                            PP        PREP      "of"
                                                      NP        NOUN*       "text"
                                                      NOUN*     "content"

        PUNC    "."
P-METRIC = 0.434341
```

```
3-2.6 (NORMAL)
8 redundancies
1 text compression
1 text sizes
4 text content loss
```

```
PARSE TREE 2 NOT DISPLAYED.
```

Perfect Syntactic Output

Fig. 1

```
CMPD    DECL     NP       NP        NOUN*      "today"
                          AJP       ADJ*       "large"
                          NP        NOUN*      "disk"
                          NOUN*     "arrays"
                 VERB*    "are"
                 AJP      AVP       ADV*       "usually"
                          ADJ*      "available"
        CONJ*    ", but"
        DECL     NP       AJP       ADJ*       "using"
                          AJP       ADJ*       "short"
                          NP        NP         NOUN*      "texts"
                                    CONJ*      "and"
                                    NP         NOUN*      "small"
                          NOUN      "dictionary"
                 VP       VERB*     "sizes"
                          NP        NOUN*      "saves"
                                    PRPRTCL    VERB*      "processing"
                                               NP         NOUN*         "time"
                                                          PP            PREP
                                                                        NOUN*
                                                                        PP

                 CONJ*    "and"
                 VP       AVP       ADV*       "still"
                          VERB*     "remains"
                          AJP       ADJ*       "attractive"

        PUNC     ","
P-METRIC = 1.406222
```

```
3-4.8(NORMAL)
1 today disk arrays
3 using text dictionary
3 using small dictionary
8 saves
8 addition
1 storage space
```

```
PARSE TREE 2 NOT DISPLAYED.
```

Syntactic Analysis Example

Fig. 2

N

natural-language/2/1/1.
natural-language/3/1/4.
natural-language representation/3/1/4.
converting natural-language text representations/2/1/1.
next frequent word/9/2/15.

O

word occurrence observations/11/1/20.
occurrences/6/2/3.
occurrences/7/4/8.
occurrences/18/3/38.
occurrences/18/2/37.
occurrences/18/1/36.
letter occurrences/6/3/4.
word occurrences/7/4/8.
word occurrences/8/2/11.
word occurrences/13/2/27.
word occurrences/19/2/44.
total word occurrences/10/1/16.
once/18/5/40.
once/15/1/32.
frequency order/10/1/16.
frequency order/9/2/15.
frequency order/9/2/15.
decreasing frequency order/8/3/12.
rank orders/9/2/15.
ordinary text/7/5/9.
ordinary text/18/5/40.
ordinary English text/8/2/11.

P

probability/10/1/16.
probabilistic terms/10/1/16.
actual proportion/18/3/38.
communications psycholinguistics/9/1/14.

Excerpt from Raw Phrase Index

Fig. 3.