AN INVESTIGATION OF THE LINGUISTIC CHARACTERISTICS
OF
JAPANESE INFORMATION RETRIEVAL

A Dissertation Presented

by

HIDEO FUJII

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 1998

Department of Computer Science

AN INVESTIGATION OF THE LINGUISTIC CHARACTERISTICS OF JAPANESE INFORMATION RETRIEVAL

A Dissertation Presented

by

HIDEO FUJII

Approved as to style and content by:

W. Bruce Croft, Chair

David W. Stemple, Member

James P. Callan, Member

Chisato Kitagawa, Member

F. Roger Higgins, Member

David W. Stemple, Chair
Department of Computer Science

To Papa Korfhage

ACKNOWLEDGEMENTS

My gratitude first goes to Bruce Croft. I am most thankful for his invaluable support throughout my academic years at the University of Massachusetts, Amherst. His advice was insightful, and I was always encouraged to challenge the deeper exploration. I thank David Stemple for his valuable comments and encouragement. I am very grateful to Jamie Callan. He always showed me his willingness to support and encourage my research. I thank Chisato Kitagawa. Theoretical background of this dissertation benefited from frequent discussions with him. I thank Roger Higgins. It was always delightful to listen to or read his knowledgeable and insightful linguistic comments.

Thank you, Sonia, for your patience - an invisible, but priceless support. Thank you, Meli and Vivi. You may or may not know, but your presence contributed greatly to this work. My last and highest gratitude goes to my parents, Terumi and Tomoko Fujii, and to my grandmother, Asae Saito, who died when this dissertation was still in progress.

This work was supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623.

ABSTRACT

AN INVESTIGATION OF THE LINGUISTIC CHARACTERISTICS OF JAPANESE INFORMATION RETRIEVAL

FEBRUARY 1998

HIDEO FUJII, B.S., KWANSEI GAKUIN UNIVERSITY
M.S., UNIVERSITY OF PITTSBURGH
Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST
Directed by: Professor W. Bruce Croft

This dissertation examines and demonstrates the effective use of linguistic knowledge in information retrieval (IR) technology. This linguistic IR research has a long history of serious but unfortunately often unsuccessful endeavors, but our retrieval experiments generally confirmed a significant performance improvement by these linguistic techniques. These experiments were realized by using a Japanese corpus. Thus, this research also serves as a case study of "linguistic information retrieval" for Japanese, as opposed to English which has traditionally been the predominant language of study.

The methodology which was taken in this study is called *grammatical paraphrasing paradigm* for the query formulation to translate a formal grammatical relationship into a retrieval strategy. To realize this paradigm, based on the theory of generative grammar, we developed a class of query strategies to be applied to a sentence in a base query having various valency structures such as transitivity or intransitivity in lexicon, or causativization

or passivization in syntax. We call this class of strategies *valency control strategies*. The most distinctive advantage of this method is the capability to draw two contingent sets of dichotomous views. The first view is the valency dichotomy that reveals the difference in strategic gain between the monovalent (i.e., intransitive and passive) and bivalent (i.e., transitive and causative) strategies. The second view is the dichotomy within a system of linguistic components, where lexical and syntactical modules have separate retrieval mechanisms.

After developing the general framework of valency control strategies from a linguistic background, especially involving the phenomenon of transitivity alternations which exist extensively in Japanese, we examined its effectiveness in a series of experiments. We found the following three uniquely important results. First, the overall result showed that most valency control query strategies considerably improved the precision. This means that linguistic knowledge is a highly valuable knowledge source in information retrieval. Second, in the valency dichotomy, the bivalent strategy improved the performance, but the monovalent method degraded it. This result indicates the usefulness of formally definable grammatical strategies in information retrieval. Third, in the linguistic module dichotomy, despite the conventional wisdom which emphasizes the local morpho-lexical information, the syntactical method was effective as well as the lexical method.

Two additional experiments on potentialization and verbal nouns were carried out, as well. The potential query strategy on verbs, which does not change the valency, showed a moderate performance improvement between bivalent and monovalent. The performance of verbal noun strategies was not as encouraging as that of verb strategies. The genitive verbal noun strategy showed a particularly clear degradation, which is probably a reflection of past data in literature showing that phrase recognition achieved only limited retrieval improvement.

Finally, this research also has a strong practical implication. We had two sets of experiments - one the relevance feedback method, the other the automatic query generation method. Our results showed that the automatic method works roughly as well as the relevance feedback method. This suggests that our method has significant practical applications because it does not rely on relevance information to improve the query performance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Because beyond this cipher's cross-barred keep
He'd see the world in all its aimless passion,
Diminished, dwarfed, and spellbound
in the symbols.
H. Hesse, Das Glasperlenspiel (1943)
trans. T. Ziolkowski

CHAPTER I
INTRODUCTION

A. Objective and Framework of This Study

This dissertation deals with the usefulness of the linguistic knowledge in information retrieval (IR) technology. More than ten years ago, Salton (1983) wrote:

> The role of linguistic methodologies in general and of syntactic analysis in particular is still unresolved for information retrieval. Before reaching a final conclusion in this area, it is wise to wait for the appearance of more sophisticated language analysis methods that are at the same time sufficiently efficient to permit incorporation into operational retrieval frameworks. Such methods should then be thoroughly evaluated to determine their actual value in information retrieval. (p.287)

This situation has, in my judgment, fundamentally not been changed, and this general conception is exactly what we want to challenge in this dissertation. Furthermore, until very recently, IR research has been focused predominantly on English text and has given less attention to other languages. Consequently, it was inevitable that a broad spectrum of linguistic properties of the English language, such as the grammar (e.g., syntax), pragmatic knowledge (e.g., characteristics of the discourse structure) and writing system (e.g., its orthographic use of the alphabets) have been assumed qualitatively and quantitatively as given research preconditions. On the other hand, the application of traditional natural language processing (NLP) technology to information retrieval has involved a limited coverage of linguistic features such as suffix stemming in morphology or noun phrase indexing in syntax. Experimental results using such techniques often showed only limited performance improvement.

Given the two factors above, i.e., the predominance of English and the limitation of language processing techniques, we often find the following criticism: Assuming that we observed a significant performance increase or decrease using some technique in English IR experiments, we do not then have a proper way to determine how the language factor (in this case, of English) influenced (or contributed) to the low (or high) effectiveness. In our view, although certain aspects of a specific language, which are measurable, may prevent the effective use of linguistic knowledge, some universally applicable methods, which also should be measurable, exist. We will discuss the details of this point in the next section.

The basic endeavor of this dissertation is to restate the IR problem in a different way, avoiding assumptions of traditional English information retrieval: First, we use Japanese text as a specific non-English medium for our experiments. As we reveal through the experiments in this dissertation, the choice of the Japanese language yields to "more sophisticated language analysis methods that are at the same time sufficiently efficient" in Salton's statement. Japanese language provides a good test-bed to prove the usefulness of linguistic knowledge in information retrieval. Second, departing from the framework of established NLP technologies, we pursue a more general linguistic approach. Thus, we conduct the retrieval experiments with some linguistically controlled specification, and if performance is in accord with the analysis of certain general aspects of the language phenomenon, we then consider this retrieval method to have the potential to be universal as an IR methodology. But, if the data do not match the analysis, we can either reexamine the underlying general language model itself, or regard the divergence as the result of the specific characteristics of the language (or a language class including the language), in this case Japanese. Therefore, our attempt is not to add Japanese to the language repertoire of IR, but instead to provide the methodology and the basis of general "linguistic information retrieval."

B. Language Universality and Language Evidence View of IR

To achieve this study's objective and framework as a linguistic approach for information retrieval, toward developing a universal methodology, first we have to elaborate the concept of universality in IR. If the performance of a linguistic module in an IR system is not affected by the choice of language, e.g., English or Japanese, the system can be called *language independent*. Only when the independence is supported by some linguistically universal properties, can the characteristic be called *linguistically universal*. On the other hand, if the performance of a linguistic module depends on some specific language aspects, the module should be called *language specific*. A language specific module may be realized by *parametrizing* the system to derive the language specific functions from a general configuration. A parameter which is related to some linguistic features has a particular value in particular languages. In other words, it sets equivalent classes of languages, and this property guarantees a unified handling of languages throughout the classes. Thus, language specificity could be considered an integral part of language universality if we regard the parameters themselves as universal properties.

As we will discuss in detail in the next chapter, this general view of technical applicability in IR corresponds to the theoretical framework of linguistics, specifically in this dissertation to the basic notions of generative grammar. In Generative Grammar, especially in the Principles and Parameters framework, aspects of a language are shaped by both the instantiation of various parameters and by generally applicable principles; furthermore, the integrated architecture of parameters and principles can be regarded as a realization of Universal Grammar (UG).

Our experiments were implemented by a retrieval system called INQUERY (Callan, Croft & Harding, 1992) which is based on the framework of a probabilistic inference network retrieval model (Turtle, 1991, and Turtle & Croft, 1991). We already knew from previous studies that INQUERY is effective not only for English, but also for Japanese, Spanish, and Chinese. At this point, a questions arises: Where did INQUERY's cross-linguistic applicability come from? Obviously, the inference network is not a model based on "linguistic" universals, but for the following reason, should be regarded as based on "human communication universals," or in more cognitive terms, a *Processing Universal* in Gass' (1989) framework. Human communication universals are realized in and observed in the "use" (or "performance," as versus Chomsky's "competence") of the language. For example, although INQUERY has very sophisticated functionality as an inference network retrieval engine, the relevance prediction of a document to the user's information need relies on the individual estimation of the relevance probability of each indexed term. For this very basic notion of relevance estimation, TF*IDF (TF: Term Frequency; IDF: Inverted Document Frequency) is the most widely used measurement. (For a summary of various variations of TF*IDF measures, see Frakes & Baeza-Yates, 1992, pp. 367-376.) Thus, TF*IDF could be regarded as a universal in human communication, as Sparck-Jones (1972) first derived it from Zipf's law of term usage in a document database. Therefore, this measure is definitely not a "linguistic universal" in the strict sense of "linguistic." Because of the non-linguistic characteristic of TF*IDF, the INQUERY engine is able to work reasonably well for any language (at least to date) as a linguistically independent retrieval system. In other words, since the INQUERY retrieval model is linguistically independent, any linguistic module, either linguistically universal or parametrized, attached to the retrieval engine can work without causing an unfavorable interaction with the retrieval engine. We might imagine a retrieval system in which the first several words of every sentence are automatically treated and retrieved specially as likely to be the subject or topic of the sentence. In such an imaginary system, texts in a verb-first (i.e., VSO (=verb-subject-object, e.g., in Irish) or VOS (=verb-object-subject, e.g., in Thai) is the basic word order of a sentence) or pro-drop (i.e., subjectless sentences are allowed) language where the subject nominal may not come at the beginning of the sentence will be mishandled. Thus, the core retrieval system and the linguistic modules must have clear modular separation and a proper functional interface.

Adopting INQUERY for the sake of working human communication universals, then how can we realize linguistic universalities and parameters in a retrieval system to make the retrieval more effective? The key to this question should be looked for in both the indexing and the query processing executed in the system. To answer this question, let us look at some basic concepts of the INQUERY system and their implication.

The inference network retrieval model is a version of the Bayesian probabilistic inference network, and is specially organized for the purpose of document retrieval. The network is divided into two subnetworks: a document subnetwork which is built at the indexing time of documents (i.e., pre-coordination), and a query subnetwork which is created at the run-time of each given query (i.e., post-coordination). In the document subnetwork, each document (represented by a document representation "node") shares one or more "concepts" (represented by document concept nodes which correspond to indexed terms) analyzed and extracted from documents, with other documents. The query subnetwork also has two types of node sets: a single query node which represents the user's information need, and the query concept nodes which are expressed in the query. Given a document database and a query, the document and query subnetworks are integrated into a single inference network at the time of the execution of retrieval. The process of retrieval is regarded as estimating a conditional probability of satisfying the user's information need (I) with a given document (D), i.e., $P(I|D)$.

As a natural extension of traditional retrieval architecture where every term (word or phrase) is simply and directly mapped to an indexing representation, we can add any analytical cue to enhance the representational capability of the text and query and to increase the (probabilistic) belief of relevance. Thus, under this network architecture, various linguistic features can be encoded and indexed as a part of "multiple document representation"

(Turtle, 1991, p. 138). Indeed, as described in Chapter II, we will construct several models that engage various kinds of linguistic information as a part of the document representation. Linguistic information becomes a knowledge source for the information retrieval.

Next, let us consider the query processing. An inference network is naturally capable of constructing a query formula using various operators (e.g., #sum, #and, #phrase, #synonym, etc.) by defining the functional and structural relationships among some specific query nodes (i.e., probabilities at input query nodes as for independent variables, and a computed value at the new node as for a dependent variable). Compared to the Bayesian inference network system, the Boolean retrieval model's operators (e.g., AND, OR, etc.) are considerably less flexible with their limitation (both in functionality and usability) of the "set operations." Moreover, in another advanced retrieval model, the vector-space model (Salton, 1971) generally lacks the capability of structural query construction because of its fundamental representational scheme of uniform arrangement of indexed terms in a multi-dimensional space.

Finally, the characteristics of inference networks can be viewed as a good example of the "language evidence" paradigm. This means language techniques should provide various kinds of evidence to improve retrieval performance. It should be noted that this language evidence paradigm contrasts with the traditional "language understanding" view of (semantics-driven) natural language processing techniques, which intends to determine the semantic contents of the text by analysis as typically seen in a question-answering system. Salton (1983) "rejects the notion that information retrieval is simply an early stage of more refined question answering" (p.267).

INQUERY shares this language evidence view by representing linguistic evidences in a multiple document representation, and by the capability of constructing a flexible query structure with various operators.

To fully utilize language evidence, the source text must be indexed with various cues to sufficiently preserve the linguistic information in the text, and a query should be optimally formulated to achieve its best potential. This optimization of the query structure is a matter of the "strategic" application of how we build the query, because the true optimal construction rests in the myriad of cause-and-effect relations of all possible query formulae and text indexing strategies. Such causal mechanisms are beyond our understanding, as we know from the significant role of heuristic knowledge in artificial intelligence (AI) applications. Good indexing is just a necessary condition to get good retrieval performance. For example, inadequate formulation of the query will produce poor effectiveness even if the text was indexed by a highly sophisticated technique. In contrast, optimizing query processing is a sufficient condition for best performance, at least under the condition of a given indexing representation of documents. Therefore, this study emphasizes the development of effective query strategies rather than exploring the satisfactory representation power of indexing.

At the end of this section, let us mention some practical issues. The notion of universal applicability of a retrieval technique which has just been discussed may sound somehow idealistic because there are many practical limitations of an implemented retrieval system. However, these problems can be often best understood and solved when they are addressed in the light of a theoretical model. For example, choosing a right retrieval operator from the available set in the IR system to achieve good performance is a highly system-dependent problem, and it also depends on the underlying retrieval model on which the operator is based. Worse, the query operators in most working IR systems are not primarily intended or designed to straightforwardly express linguistic relationships. So, the effective use of retrieval operators is based on practical knowledge, while at the same time its effectiveness depends also on which kind of linguistic phenomenon we are applying it to. In other words, there is a right operator for a right linguistic opportunity. For example, the effectiveness of various operators on simple compounds was reported for Japanese (Fujii & Croft, 1994) as well as for English (Croft, Turtle & Lewis, 1991). Indirectness and limited scope of universal applicability of linguistic knowledge in IR is an inherent problem of *any* application system (including even human intelligence itself) in the real world.

C. Review of Previous Research

There are two sub-areas in language-oriented IR research which fall within the purview of this dissertation. One is the exploration of general linguistic applicability and characterization of various languages in IR. The other is the development of the art of query formulation to represent certain effective linguistic aspects.

First, as for the matter of linguistic applicability in IR, in my view the general research method has not yet been systematically explored, despite there being a number of studies of non-English retrieval even from the early period of development of IR systems, such as Salton's SMART system (1971) for English, German and French, or Pevzner's Empty-Nonempty-2 system (1972) for Russian and English. To date many systems have been developed, among others, Hebrew (Attar, et al. 1978; Choueka 1990), Japanese (Fujii and Croft, 1993), Chinese (Nie, Brisebois and Ren, 1996), Korean (Lee and Ahn, 1996), Arabic (Wien, 1996), and Spanish (Ballesteros and Croft, 1996). However, these studies tended either to focus the idiosyncratic characteristics of a specific target language (often in ad hoc manners), or to ignoring the effecting factor of the language(s) used in the experiments. Furthermore, syntactical modeling of IR has been especially overlooked (as Salton stated in the beginning of this chapter) in most cases. In contrast to these studies, our approach has more direct resonance with the following statement about the distinct effects of word distance in Japanese and English queries:

> "The slower increase [of precision value along the change of proximity window size] in Japanese suggest the strong locality of word distribution in the text. ... we predict that syntactic approach (of a sentence) in Japanese will be more effective than in English." (Fujii & Croft, 1994, p.93)

Thus, in this view certain linguistic (especially syntactic) universal IR methodology exists, however specific characteristics of a language determine the degree of its tractability for retrieval, and may help (in Japanese) or hinder (in English) the universality from being revealed. Ultimately, the point implicitly urging here is the level of adequacy of methodology applicability in linguistic IR.

Furthermore, in the recent rise of multilingual information retrieval, there has been one commonly observed methodological aspect across all the studies. That is, they tend to utilize conceptual equivalency to achieve cross-lingual capability by such means as lexical semantic relations in dictionaries (Hull and Grefenstette, 1996), alignment of parallel texts (Davis, 1996), machine translation (Gachot, Lange and Yang, 1996), etc. Thus, the basis of these studies is not the formal relationship of grammar, but the use of communication-based semantics, in other words, human communication universals rather than linguistic universals. Again, the target of these studies and the target of this study differ in terms of the domain of linguistic discipline and the class of universality.

Next, concerning the methodology of query formulation to effectively represent certain linguistic terms and relationships, let us consider the example of a Boolean retrieval system. In such a system, it is a popular practice to use operators, e.g., AND to restrict (or OR to expand) query semantics or "proximity" for idiomatic phrases or compounds. The major problem here is that such query construction is based either on the expertise of the manual translation from the query description to the formula (e.g., Gachot, Lange, and Yang, 1996, among others), or on the automatic process only for a limited domain of linguistic phenomena such as noun phrase recognition in syntax (e.g., Fagan, 1992; Fujii and Croft, 1993, among others) or query expansion in lexical semantics (e.g., Voorhees, 1994; Xu and Croft, 1996, among others). Furthermore, as we already described in the last section, there is the problem that retrieval operators in most IR systems are not optimally designed to express linguistic structural properties. Thus, IR methods and architecture for linguistic query formulation have not been established to accomplish "sophisticated language analysis methods ... to permit incorporation into operational retrieval frameworks" as in Salton's words cited in the introduction of this dissertation. Therefore, the language evidence approach of the inference network model should be tested in our experiments to prove its technical accountability for effective linguistic query formulation.

D. Presentation of Sections

The following chapters are divided into three major parts: Chapter II is an attempt to model the use of linguistic knowledge to form an effective query strategy. Chapter III describes the methodology we utilized as the basis of our experiments. Chapter IV is a report of a series of experiments to test the model described in Chapter II and the hypotheses stated in Chapter III.

The major part of Chapter II concerns the grammatical model of query formulation strategy. The so-called "grammatical paraphrase model" (in contrast to the "arbitrary semantic paraphrase model") is described. Since grammaticality becomes an issue in this model, the essential concepts of the Principles and Parameter Approach of generative grammar will first be described. The central argument here is that careful and sophisticated linguistic analysis of a language, in this case Japanese, can shape a query strategy. As a prominent test case of such an endeavor, a strategy based on the control of the number of verbal arguments, namely the "valency control strategy" in syntax (e.g., passivization or causativization) or in morphology/lexicon (e.g., intransitivization or transitivization), will be defined, followed by the theoretical description of transitivity alternation in Japanese. Thus, in this experimental scheme we will be able to test the retrieval performance both in the valency dichotomy (i.e., monovalency vs. bivalency, in other words, passive vs. causative, or intransitive vs. transitive), and in the linguistic module dichotomy (e.g., syntax vs. lexicon).

Chapter III is about the methodology of our experimentation. After describing the hypotheses derived from Chapter II's model, we will develop suitable methods of text indexing and query formulation to encode the linguistic knowledge in both the text database and the query.

In Chapter IV, a series of retrieval experiments are conducted to answer the three major research questions of this dissertation: first, whether a general improvement in grammatical paraphrasing (using the valency control method) is possible, or not; second, whether the bivalency strategy is more effective in retrieval than the monovalency strategy (or vice versa), or not, in the valency dichotomy; and third, whether a lexical method performs better than a syntactical method (or vice versa), or not, in the linguistic module dichotomy. This framework of examination substantiates the basic hypotheses and the design of the experiments described in Chapter III.

Chapter V, the last chapter of this dissertation, is the conclusion, and clarifies what directions might be of value and what might not, in future research.

E. Research Contributions

A general and important contribution of this dissertation is to exhibit the substantial role of general linguistic knowledge in information retrieval, both by proposing a theoretical model of grammatical query formulation and by demonstrating its actual performance improvement through a series of retrieval experiments. So, we generally regard this study as an investigation of "linguistic retrieval." Careful and sensitive linguistic analysis, specific or general, is necessary to lead to solid results in linguistic retrieval. We chose the Japanese language as a test-bed to give our theoretical model a grounded, real world meaning. Japanese was a conveniently qualified language for that purpose. Thus, this research is an investigation of linguistic information retrieval through any means. Our retrieval model can be regarded as just a first step toward a "universal query strategy" through analysis formulated within a generative grammatical framework. A grammatical paraphrase query model such as the valency strategy we propose in this dissertation is such an endeavor. The research direction of this work has rarely been pursued in traditional IR research.

The above contribution became possible because of the adoption of the language evidence paradigm in retrieval. The inference network retrieval system adopted in this study is a one of a kind system that supports this approach. This investigation should be a highly beneficial demonstration of the construction of an information retrieval system in which linguistic properties are deeply involved.

CHAPTER II

A MODEL OF LINGUISTIC KNOWLEDGE IN INFORMATION RETRIEVAL

A. Introduction: Linguistic Knowledge and IR

We start this chapter by asking how language processing contributes to IR technology in general. The use of natural language processing techniques for IR has a long history, but one strewn with constant debate: whether sophisticated language processing may or may not pay off in improved retrieval performance. Other natural language processing technologies such as machine translation usually process only a relatively small amount of text (e.g., tens or hundreds of sentences), or require large amount of data only at the training stage, so the system can spend as much time as needed to learn the characteristics of the training data, but then the actual performance would proceed on only a small input text. The problem is that they are usually too slow to process a great number of texts (e.g., thousands to millions of documents) which is the practical size for retrieval. More seriously, language analyses have often been questioned as to significant improvement in effectiveness. We see such cases in literature as the identifying of noun phrases (Fagan, 1987; Krovetz, 1995; among others), or the adding of parts-of-speech information (Krovetz, 1995). Linguistic approaches were, thus, considered as too "expensive" to proceed with, and the results as "disappointing" (Salton, 1983, p.286). However, this pessimistic situation could change using the "language evidence" approach differently, as we convinced ourselves through the development of theoretical models and the examination of our experimentation in this study.

Throughout this dissertation, as we explained in section I.B, we emphasize formulating a query strategy rather than developing an indexing method. Now, let us consider a linguistic framework of the generative approach as a developmental framework for query formulation methodology. The generative approach searches for formal descriptions of linguistic structures and operations in *Universal Grammar* (UG), such as hypothesizing the transformation of an underlying syntactic structure to a surface sentence expression. This hypothesis must be justified by checking the grammaticality of the produced sentence by the "data" - a native speaker's judgment. So, this is a fairly rationalistic deductive approach to "explain" human language competence.

Generative grammar plays a straightforward role in building a query strategy. Thus, we can specify various grammatical properties of a sentence (or even of a smaller unit like a word, if applicable), say, passivization for example, to encode the retrieval cues systematically in a query. The advantage of generative grammar is that we can always have reference to the original underlying structure as the basis of the encoding. For example, the underlying structure of a passive sentence is essentially shared with its corresponding active sentence. Although it is important to know the gaps between grammatical theory and information retrieval, there is a compromise view. We know that the objective of generative grammar is to understand human linguistic competence, which covers an infinite number of "grammatical" sentences - quite different from the practical goal of information retrieval. However, it is possible to apply the generative framework to IR in a different way. Assuming a certain underlying structure and limited range of practically meaningful derivations[1] of the structure, we concentrate on the variations of derived expressions and their relationships at the surface level, knowing that such derivations are conditioned by "principles and parameters" of generative grammar. This *Principles and Parameters Approach* (PPA) must be suitable for query strategy building because such parameters and principles often show a fine granularity in "controlling" the production of forms.

Despite attempts even from the early stage of computer application development in the 1960s, the relationship of generative (transformational) grammar and IR has remained at a very conceptual or basic level, such as the use of phrase-structure rules to extract noun phrases in a document. Furthermore and unfortunately, there was a historical trend in IR to deemphasize language analysis, as we already described in the beginning of this section. It was hardly conceivable that a grammatical analysis such as the parsing of all sentences in documents could directly contribute to performance improvement. Application of generative grammar to IR remains an almost untouched area even today.

The Principles and Parameters Approach described above and the *Government and Binding* (GB) approach are used almost interchangeably in the theory of generative grammar. In the GB framework, a principle can be parametrized to explain and generalize its applicability to any language for the sake of Universal Grammar. The GB Theory is usually divided into seven sub-theories (or *modules*): 1) X-Bar Theory, 2) Theta-Theory, 3) Government Theory, 4) Case Theory, 5) Binding Theory, 6) Bounding Theory, and 7) Control Theory. They are briefly characterized as follows:

[1] We put on this reservation because a natural language grammar itself is as powerful as recursively enumerable, but there are cognitive bounds and practicality factors for IR.

1. **X-Bar Theory** provides the derivation of the phrase structure of a sentence.
2. **Theta Theory** deals with the assignment of thematic roles to noun phrases.
3. **Case Theory** deals with the providing of nominals with some 'Cases' such as subjective and objective.
4. **Government Theory** defines a formal relationship, "government" between a constituent ("governor") and certain nodes in a constituent structure.
5. **Binding Theory** deals with how to describe the determination of the relationships of pronominals or anaphors to their antecedents.
6. **Bounding Theory** sets limits on how distant a movement is allowed to be in the phrase structure.
7. **Control Theory** deals with the interpretation of VP with no overt subject.

Considering the applicability of these modules to IR, in this study we will mainly concern ourselves with the first three theories, i.e., X-bar theory, Theta theory, and Case theory, and will seldom address the Binding, Bounding and Control theories. The Government theory will be treated less explicitly. However, it does serve as a fundamental mechanism for various principles in the first three modules. Because they are outside of the scope of this paper, readers should consult related publications for detailed discussions of these four theories.

B. Grammatical Paraphrase Model of Query Formulation

The *grammatical paraphrase model* is a general methodology for developing a query formulation strategy. Similar to the ordinary query expansion/modification method using the synonymy relations in a thesaurus, a class of semantically (quasi-)equivalent correspondence of expressions (sentences, phrases, etc.) can be used in a query as paraphrasing. However, unlike many thesaurus methods, which often improve recall principally by extending the keyword coverage in the query (i.e., which work as a recall enhancement device), the main point of paraphrasing is to find a better way to express the query context, and consequently, it is suitable to attaining higher precision (i.e., it works as a precision enhancement device). The essence of the grammatical paraphrase model is to paraphrase the original query based solely on grammatical relationships - the equivalence of semantics does not matter to the operation, but may result from the operation, if it so happens. So, it is a non-empirical method, and can be contrasted to semantics oriented methods such as "concept-based paraphrasing" or "arbitrary semantic paraphrasing." In other words, grammatical paraphrasing intends to be more formal, well-defined, and potentially brings more automaticity to the linguistic retrieval process than semantics oriented paraphrasing. A sentence is not the only material subject to grammatical paraphrasing, but any smaller unit such as a phrase, a compound, or even a simple word may be the subject of the operation, if we can find a grammatical property in such a unit. Therefore, the relationship may be found not only from syntax, but possibly from morphology or lexicon. Note that grammatically paraphrased expressions *may* share an identical underlying (D-)structure in order to preserve the semantics we may have wanted, but it is not always necessary. The grammatical paraphrasing can be established based on the formal "correspondence" between two expressions rather than their "deep" "equivalence." In the next subsection, a group of so-called "valency control strategies" that will be the primary focus in our experimentation in Chapter IV, will be defined.

C. Valency Control Strategies

Valency control strategy (or "valency strategy" for short) is an example of the grammatical paraphrasing model of query formulation described at the beginning of this section. The domain of the valency control operation does not have to be fully configured as a complete sentence, but may be a limited part of the structure if it is justified as an operational model that gives a better retrieval performance in practice. Thus, setting the applicable domain or context is crucial for any strategy to achieve a good result, and the valency strategy is no exception. The valency control strategy defined in this study operated exclusively on a verb and its obligatory arguments in the original query sentence, but excludes optional adjunct elements as factors varying from sentence to sentence in applicable texts. It formulates a new sentence pattern by increasing or decreasing the valency from that of the original sentence. Let us elaborate the meaning of these operational characteristics of the valency control strategy as follows.

There is a well-recognized view that the (main) verb has a pivotal role in forming the basic structure of a sentence. Thus, the verb serves the role of lexical head of the sentence, and provides its argument structure as a basic building block for the framework of the sentence. As a verb's inherent property, an argument structure determines the number of structural positions (i.e., *valency*) and the thematic roles (or *theta-roles*) of arguments. Here, an argument is defined as an obligatory noun phrase (NP) which co-occurs with the verb to make the verbal structure well-formed. For example, a verb "*break*"(trans.) has the following argument structure:

BREAK: (AGENT <PATIENT>)

In this case, the verb BREAK has two arguments which are bound to specific theta-roles, AGENT (i.e., doer of the action) and PATIENT (i.e., an entity which undergoes the action). AGENT is specified at the external position of argument structure (i.e., as an *external argument*), and PATIENT at the internal position (i.e., as an *internal argument*). A ditransitive verb has one external argument and two internal arguments. For example, "*give*" has AGENT (or SOURCE) as an external argument, and THEME (i.e., an entity which is taken in the event) and GOAL as internal arguments.

As we defined above, valency is defined as the number of arguments which a verb takes obligatorily in a sentence. At S-structure, an intransitive verb (V_i) has a valency of one (i.e., it is *monovalent*) of the subject, a transitive verb (V_t) a valency of two (i.e., *bivalent*) of the subject and the (direct) object, and a ditransitive verb a valency of three (i.e., *trivalent*) of the subject and direct and indirect objects. Since valency comes from the argument structure stored in the lexicon, from where it is introduced into D-structure, valency is persistent throughout the path from lexicon to D-structure, to S-structure, as required by the *Projection Principle*. Thus, as far as valency status is concerned, valency is a highly robust property by which to characterize a sentence, and it makes sense to utilize arguments, excluding adjunct elements, as the basis of query formulation. However, it should be noted that the above assertion does not guarantee a decisive tie between the subject of the sentence and the external argument, or between the object of the sentence and the internal argument at D-structure, although this does happen in many instances. One statement of the form of such a relationship is called the *Unaccusative Hypothesis*, and will be discussed later in this section.

Corresponding to the linguistic components where a strategy operates effectively, there are two ways of controlling valency: syntactic and lexical. The *syntactic valency strategy* uses syntactic means such as arranging the words, introducing auxiliary elements, etc. to change the valency of a verb. *Passivization* and *causativization* are typical strategies of this kind. Thus, a query sentence can be passivized or causativized to be syntactically paraphrased. Passivization changes the voice and decreases the valence by one from the original sentence, as the object in the underlying D-structure moves to the subject position in S-structure[2], and the original D-structure subject is either simply removed from the sentence or expressed as an optional adjunct phrase. On the other hand, causativization increases the valency by introducing an extra argument, i.e., a causer at the subject position, so that the original subject becomes the object of the verb.

Lexical valency strategy is another type of valency strategy which modifies the verb form itself at the lexical level to control its valency. Categories of verb valency are *intransitivity* (valency=1), *transitivity* (valency=2), and *ditransitivity* (valency=3). For example, the Japanese verb root "MI" derives three distinct verbs, MI-e (be.visible), MI-ϕ (see), and MI-se (show), respectively. Thus, in this example, they are morphologically related

[2] It is theoretically possible that query modification by passivization involves further process after S-structure, such as setting up the quantifier representation. However, practically speaking in IR, this is usually just a less significant problem because such fine grain features usually do not come up as index terms. Passivization with an intransitive verb is another problematic case. In English, there are so-called *pseudo-passive* or *prepositional passive* sentences (see Levin & Rappaport, 1995, p.144), e.g., "This bed was slept in by President Lincoln." In Japanese, so-called indirect passive sentences (see Kageyama, 1993, p.60) are possible, e.g., John-ga Petto-ni SIN-are-ta (John-NOM pet-by die-PASS-PAST; 'John was disturbed by his pet's dying' or 'John's pet died on him.'). In these cases, as far as original main verbs are concerned, a new argument nominal (similar to causativization) is introduced, therefore the valency is increased rather than decreased. These passivization of intransitive sentences are analyzed in terms of their underlying structures. See the discussion about the Unaccusative Hypothesis in section E of this chapter.

by sharing an identical root, but with distinct suffixes. They are instances of derivational morphology, and share certain semantics and form characteristically expressed by the root, though each item (as a word) may display its own idiosyncrasies as a lexicalized item.

It is important to distinguish the above syntax-lexicon dichotomy from the sentence-word dichotomy. Although the sentence is the domain of syntactical construction, and lexical items are in essence words, the inverses of these propositions do not necessarily hold. For example, there are conceivable cases (at least in Japanese) where some syntactic operation is applied on some word construction as a morphological phenomena. According to Kageyama's (1993) modular theory of morphology, morphology is a linguistic component for word formation which has interfaces both to syntax and to lexicon, and he demonstrated that some Japanese compounds have syntactic nature. We may sometimes call a valency strategy a *morphological strategy*, if we characterize the strategy as operating on word formation, which is possibly either lexical or syntactical.

As examples of the above discussion, let us examine the following English sentences (1) and their Japanese translations (2) :

(1) a. The tree fell down.
  b. A woodcutter felled the tree.
  c. The tree was felled (by the woodcutter).
  d. A woodcutter let the tree fall down.

(2) a. Sono Ki-ga TAO-re-ta.
  [that tree-NOM fall-INTR-PAST]
  b. Kikori-ga sono Ki-o TAO-s-ta. (s-t=>sit)
  [woodcutter-NOM that tree-ACC fall-TRANS-PAST]

c. Sono Ki-ga [Kikori-niyotte] TAO-s-are-ta.
[that tree-NOM [woodcutter-Caused.by] fall-TRANS-PASS-PAST]

d. Kikori-ga sono Ki-o TAO-re-sase-ta.
[woodcutter-NOM that tree-ACC fall-INTR-CAUS-PAST]

The formulations of these two sentence sets correspond more or less in parallel fashion, characterized as follows: In (1a, 2a), the sentences show active sentences with intransitive verbs; in (1b, 2b) the sentences show active sentences with transitive verbs; in (1c, 2c) the sentences show passive sentences with transitive verbs; and in (1d, 2d) the sentences show causative sentences with intransitive verbs.[3] Furthermore, verb pairs formed by an intransitive-transitive alternation (we call them *transitivity doublets* or simply *doublets*), i.e., (*fall,fell*) in English and (*TAO-re,TAO-s*) in Japanese are morphologically related (in English by an ablaut, and in Japanese by a suffix), and they also share the semantics of the event of "falling (of a tree)." Overt marking for transitivity doublets like (*fall,fell*) in (1) is very rare[4] in English, and zero-morphological pairs (polysemous relation in semantic terms) are more common, for example, *break, open, stand, burn*, etc. A typical example is shown in (3) (and the Japanese counterpart (4)).

(3) a. The cage broke.

b. The macaque broke the cage.

[3] The acceptability of causativization of an intransitive verb varies and depends on its semantic properties. This phenomenon is similar to the acceptability of passivization of an intransitive verb. Kageyama (1993) pointed to the "CONTROL" (ability to change the thing by one's own intention) feature in the lexical conceptual structure, in addition to aspectual meaning, as a crucial factor. However, causativization is much less constrained by this factor because it simply imposes a causer (supposedly with CONTROL) on the event. So, if we can interpret the event as being caused in some indirect way, then even if the verb lacks CONTROL, it could be acceptable, even in the case of a causative intransitive. For example, (4d) "Saru-ga Ori-o KOWA-re-sase-ta" (The macaque made the cage break(intr.)) is understandable if we have a situation such that the cage was equipped a kind of self-destruction mechanism, and the macaque triggered the switch. We will return to this topic in Section II.D and E when the issue of unaccusativity is discussed.

[4] Other examples are (*lay, lie*), (*raise, rise*), and (*seat, sit*).

c. The cage was broken (by the macaque).

d. The macaque made the cage break.

(4) a. Ori-ga KOWA-re-ta.
[cage-NOM break-INTRANS-PAST]

b. Saru-ga Ori-o KOWA-s-ta. (s-t=>sit)
[macaque-NOM cage-ACC break-TRANS-PAST]

c. Ori-ga (Saru-ni) KOWA-s-are-ta.
[cage-NOM (macaque-Caused.by) break-TRANS-PASS-PAST]

d. Saru-ga Ori-o KOWA-re-sase-ta.
[macaque-NOM cage-ACC break-INTRANS-CAUS-PAST]

No matter whether overt or covert, transitivity alternations (as in a-b) are under the control of the morphological rules of words, and they should be analyzed as lexical matters. Viewing these sentences in (1)-(4) as IR queries, (a) and (b) are examples of lexical (and morphological) query formulations, and (c) and (d) are syntactical query formulations. In terms of valency, (a) and (c) are monovalent, and (b) and (d) are bivalent. (Table 1)

Table 1. Varieties of typical valency control strategies.

valency component	Monovalent	Bivalent
Lexical	intrans+active (a)	trans+active (b)
Syntactical	trans+passive (c)	intrans+causative (d)

As we noted in the beginning of this section, these variations are paraphrasally compatible, but not strictly equivalent in various ways. First, obviously, by the passivization (c) of an active transitive sentence (b) the information about the original subject would be lost, unless the oblique NP is specified. Causative construction (d) obviously introduces an extra noun phrase as a "causer" subject into the original intransitive sentence (a). Next, while a transitivity doublet is a pair of morphologically related items, they are separate lexical items, and each may demonstrate its own idiosyncratic meaning. For example, "*fell*" as a marked[5] form of "*fall*" has a narrow meaning of 'cutting down of a tree or a similar standing object.' Every word may show polysemy, such as "to *fall* in love" (situation), or "teki-o *TAO-s*" (enemy-ACC fall-Trans; 'to defeat the enemy'). As a corollary, an active sentence with an intransitive verb (1a, 2a) and a passive sentence with a transitive verb (1c, 2c) are not the same. An intransitive verb often does not indicate the agentivity in an event - for example, a tree can "fall" by itself by no specific cause. In contrast, the transitive verb in a passive sentence still specifies explicitly or implicitly the original agent as an oblique NP brought from the external argument of D-structure even after the passivization transformation. Similarly, an active transitive sentence (1b, 2b) and a causative intransitive sentence (1d, 2d) are not semantically identical.

Regardless of these differences, a compatibility between these sentences is apparent through various combinations. For example, we recognize a parallel meaning between intrans+active (a) and trans+passive (c), also between trans+active (b) and intrans+causative (d) because they have the same valency value, in this case one and two, respectively. Furthermore, no extra meaning is superimposed on the active sentence (b) by its passivized sentence (c), and there is a transparently identical meaning in the internal argument at D-structure. On the other hand, the causation (d) simply assigns

[5] In Japanese, judgment of markedness, whether of a transitive or intransitive, is not obvious because each form often has its own suffix. Refer to the discussion in the next subsection.

an extra causer as an external position, and the description of the original event ("the falling of a tree" of (1a)) is embedded as is the internal argument of the causative verb. Thus, the original argument structure is again preserved with no information loss. Though passivization and causativization show different forms of transformation, they preserve meaning as transparently as syntactic paraphrasing, in contrast to the transitivity alternation of (a) and (b). Moreover, passivization and causativization are almost always possible when the original sentence is grammatical, as seen in (5) and (6):

(5) a. *The ball kicked.
b. The kid kicked the ball.
c. The ball was kicked (by the kid).
d. *The kid let the ball kick.

(6) a. *Booru-ga *KER-ar-ta. (r-t=>tt)
[ball-NOM kick-INTRANS-PAST]

b. Kodomo-ga Booru-o KER-ϕ-ta. (R-t=>tt)
[man-NOM rope-ACC kick-TRANS-PAST]

c. Booru-ga [Kodomo-niyotte] KER-ϕ-are-ta.
[ball-NOM [kid-Caused.by] kick-TRANS-PASS-PAST]

d. *Kodomo-ga Booru-o *KER-ar-ase-ta.
[kid-NOM ball-ACC kick-INTRANS-CAUS-PAST]

Unlike "fall" in (1), "kick" in (3) does not have the intransitivity required to make (5a) grammatical. Therefore, the causative sentence derived from it, (5d) is also ungrammatical. Only the transitive form (5b) and the passivized sentence (5c) are grammatical. Likewise in Japanese, (6a) and (6d) are ungrammatical because the transitive verbal root "KER" ('kick') does not take a suffix "-ar" (or "-e") to derive an intransitive verb.

Consequently, if we are to implement a valency control query generator, the passivized sentences (5c, 6c) can be produced automatically by paraphrasing (5b, 6b), but the production of (5a, 6a) must be blocked. To distinguish whether a verb can be intransitivized (or transitivized) or not, we always have to check the lexicon. It should be noted that such intransitivizability (or transitivizability) varies verb from verb, and also language from language. For example, English sentences in (7) are similar to (5), that is, the verb *rescue* does not have an intransitive subcategory. However, in the Japanese correspondences (8) the transitive verb TASK-e ('rescue') has the intransitive partner TASK-ar ('being resulted to be rescued')[6] (8a), and consequently the causative sentence (8d) is also grammatical.

(7) a. *The baby rescued first.
b. The fire fighter rescued the baby first.
c. The baby was first rescued (by the fire fighter).
d. *The fire fighter let the baby rescue first.

(8) a. Akanbou-ga Sakini TASK-ar-ta. (r-t=>tt)
[baby-NOM first help-INTRANS-PAST]
b. Syoubousi-ga Akanbou-o Sakini TASK-e-ta.
[fire.fighter-NOM baby-ACC first help-TRANS-PAST]
c. Akanbou-ga [Syoubousi-niyotte] Sakini TASK-e-rare-ta.
[baby-NOM [fire.fighter-by] first help-TRANS-PASS-PAST]
d. Syoubousi-ga Akanbou-o Sakini TASK-ar-ase-ta.
[fire.fighter-NOM rope-ACC first help-INTRANS-CAUS-PAST]

This kind of ungrammaticality does not only occur with transitive verbs. In (9), the English verb "occur" has only intransitivity, but the Japanese

[6] There is a different verb with a similar meaning, SUKUW-ϕ which has only transitive form, but no intransitive partner as 'rescue' does in English.

counterpart has both an intransitive form (OKO-r 'occur') and transitive form (OKO-s[7] 'cause to happen') as in (10).

(9) a. The accident occurred.
b. *The student occurred the accident.
c. *The accident was occurred (by the student).
d. The student made the accident occur.

(10) a. Jiko-ga OKO-r-ta. (r-t=>tt)
[accident-NOM occur-INTRANS-PAST]
b. Gakusei-ga Jiko-o OKO-s-ta. (s-t=>sit)
[student-NOM accident-ACC occur-TRANS-PAST]
c. Jiko-ga [Gakusei-niyotte] OKO-s-are-ta.
[accident-NOM [student-Caused.by] occur-TRANS-PASS-PAST]
d. Gakusei-ga Jiko-o OKO-r-ase-ta.
[student-NOM accident-ACC occur-INTRANS-CAUS-PAST]

Finally, it is crucial to examine the verbal morphology of a specific language to implement a valency control query system for that language. In the following subsections, we perform this task on Japanese as a case study.

As we discussed in this subsection, we have introduced three types of valency control strategies in the grammatical paraphrase query formulation model: i) lexical valency control, ii) syntactic valency control, and iii) morphological valency control. Lexical and syntactic valency controls are mutually exclusive. Morphological control, however, is not, because instances of a morphological strategy can be realized either in the lexical or the syntactic component. Thus, whether an operation is morphological or

[7] The transitive suffix "-[a/o]s" is called the short causative (Miyagawa, 1989), and is distinct from the inflectional causative suffix (i.e., long causative) "-[s]ase," though they are diachronically related. The short causative shares more or less the meaning of "causation" with the long causative, and is a very productive formation. So, to find an ungrammatical transitive example in Japanese is probably more difficult than in English. Moreover, as pointed out in Fujii and Kitagawa (1997), it should be carefully tested whether a short causative belongs to a lexical or a syntactic operation, depending on the verbal stem which is attached to the suffix. We will discuss this in more detail in the next subsection.

not will not be the central problem in our discussion (to make it more comprehensible), and we will raise issues of lexicality of words and syntacticality of phrase structures separately in the following two sections (II.D and II.E).

D. Lexical Valency Control with Transitivity Alternations

In the previous section, we introduced intransitivization and transitivization as lexical ways to control verb valency in a sentence. In those discussions, we observed in English and Japanese sentences (1)-(10) how the transitivity varies grammatically and semantically from verb to verb, and from language to language. Therefore, we argued that to understand the morphological conditions of transitivity change in a language, it is essential to implement a computational model of lexical valency control for query strategy development. There are two approaches to achieving this goal. One depends totally on empirical data such as dictionary entries. The other coordinates a theoretical framework with empirical data. The dictionary method has two immediate drawbacks: i) a dictionary often lacks consistency and coverage - some transitive forms of intransitive counterparts (or vice versa) are sometimes missing in a manually built dictionary; ii) information about the mutual relationship of the partnership of a transitivity doublet is not often explicitly defined. We argue that the second approach, that is the theorization or theoretical augmentation of experimentation, is essential to frame our agenda, to design the experimentation, and to interpret the data. For example, consider a task distinguishing an intransitive (derivational) suffix or a passive (inflectional) suffix (or an auxiliary) in order to define and justify the selection of indexing units. We may decide to discard the inflectional suffix, and to keep and mark the derivational one[8]. This is not a trivial operation, especially for a highly synthetic (i.e., inflectional or agglutinating) language where a word often has a complex construction of more than one morpheme with various grammatical functions. Japanese, a typical agglutinating language, has three[9] intransitive suffixes, "-ar," "-e," and "-i." Here, the suffix "-ar" shares some characteristics (sound, valence function, and semantics) more or less with a passive suffix "-are." Should we somehow intermingle the analysis of "-ar" and "-are"? Or, should "-ar" be treated equally with "-e" or "-i," but be separated completely from "-are"? Without a critical theoretical examination of the relationships between these suffixes, we do not have solid ground on which to base our query experimentation on whether the construction is based on words (i.e., lexicon) or phrases (i.e., syntax). To explore the phenomenon of Japanese transitivity alternations in the verbal morphology from the information retrieval point of view, we have to describe how the transitivity and intransitivity of Japanese verbs are lexically or morphologically manifested. Thus, first, we will briefly describe the overall system of Japanese predication. Then, the issue of transitivity alternation, which is the basis of lexical valency control, will be discussed. Finally, we will connect these linguistic concepts to the practical elements of indexing and query formulation in IR.

First, let us look at the system of Japanese predication. Similar to English, there are two principal types of predication: adjectival and verbal. Since Japanese adjectival expression (11a) does not involve the verbal

[8] In this sense this operation can be regarded as a generalized variation of both stemming and stopword removal by finding and removing an unnecessary element in a sentence for indexing.

[9] If we count the zero (-ϕ) ending, there are four. Moreover, there are many idiosyncratic cases.

transitivity morphology[10], we exclude it from our scope. In the Japanese verbal system there are two major encodings: verbs and verbal nouns (VN). A verbal noun behaves like a nominal element, but also has a verbal property, which is characterized by the argument structure. When used as a predicative element, it should be incorporated with a formal functional verb (i.e., *copula*) "S" (the dictionary form is "Suru")[11] (English '*do*'). Verbal nouns are mostly loan words from Chinese, but some are from English and other Western languages. In (11b), KIR-ϕ is a native Japanese verb for 'cut.' "Setudan" (11c) and "Katto" (11d, a loan from *cut*) are verbal nouns of Chinese and English origin, respectively.

(11) a. Sono-Otoko-ga Waru-i.
[Det-man-NOM bad-PRES] ('The man is bad.')

b. Sono-Otoko-ga Tuna-o KIR-ϕ-ta. (R-t=>tt)
[Det-man-NOM rope-ACC cut-TRANS-PAST] ('The man cut a rope.')

c. Sono-Otoko-ga Tuna-o Setudan-S-ta. ('S-t=>sit')
[Det-man-NOM rope-ACC cut(VN)-DO-PAST] ('The man cut a rope.')

d. Kantoku-ga Eiga-no Sono Bamen-o Katto-S-ta. ('S-t=>sit')
[director-NOM Movie's that scene-ACC cut(VN)-DO-PAST]
('The director cut the scene from the film.')

[10] Like English "-en" (e.g., "soften"), there are a few Japanese suffixes such as "-m" and "-ram" which derive a verb from an adjectival root. The distinction of transitive and intransitive is made by adding a standard verbal suffix listed in Table-3. Thus, from "-m" and "-ram," doublet patterns "-m-ar/-m-e" and "-ram-ϕ/-ram-e" are realized, respectively.

(1) John-no Kao-ga AKA-i. (John-GEN face-NOM red-Adj-PRES: 'John's face is red.')
(2) John-no Kao-ga AKA-ram-ϕ-ta. (m-t=>nd)
(John-GEN face-NOM red-INTR-PAST: 'John's face became red.')
(3) John-ga Kao-o AKA-ram-e-ta. (John-NOM face-ACC red-TR-PAST: 'John made his face become red.')

Since these verb-making suffixes have little productivity, we do not consider them a part of our valency control system. However, there is no reason to exclude their verbal forms. Thus, we regard such an original adjectival root plus a verb-making suffix as a new verbal root like AKARAM in above example. We should note that the degree of separation of adjectives from the verbal system varies language by language. In some languages, whole adjectival concepts may be represented by "adjectival verbs" so that they should be considered an indigenous part of the system of transitivity alternations.

[11] "S" conjugates irregularly. When its present infinitive attached to the present morpheme "[r]u," it becomes the standard dictionary form "Suru" (i.e., S-u=>Su-ru : DO-PRES).

In contrast to (native) verbs, there are many characterizations of verbal nouns as abstract, complex, formal, technical, etc. However, the most important aspect of the Japanese verbal noun for our purpose is that no morphological marking for valency alternation is manifested on it[12]. Thus, a verbal noun as a lexical entity has no morphological means[13] to change its argument structure, and the only possible methods are syntactical ways by passivization ("S-are" (DO-PASS)) or causativization ("S-ase" (DO-CAUS)) of the incorporated verb "S." (12a,b)

(12) a. Tuna-ga (Sono-Otoko-niyotte) Setudan-S-are-ta. ('S-t=>sit')
[rope-NOM (Det-man-Caused.by) cut(VN)-DO-PASS-PAST]
('The rope was cut (by the man).')

d. Nakama-ga John-to-Emily-o Dansu-S-ase-ta.
[friends-NOM John-and-Emily-ACC dance(VN)-DO-CAUS-PAST]
('Friends made John and Emily dance.')

On the other hand, Japanese native verbs have a complex and rich morphological system. First, verbs are divided into two groups. One is a group of non-doublet verbs, i.e., these verbs do not have an transitive or intransitive partner. The other group is doublet verbs, i.e., each verb does have its own transitive or intransitive counterpart that shares the same root. (13a) and (13b) show some examples of both groups.

(13) a. Non-doublet Verbs (trans.): NAGER-ϕ 'throw'; KER-ϕ 'kick';
NOZOK-ϕ 'peep at'; KOKOROMIR-ϕ 'try'; DAK-ϕ 'hug';
NIGIR-ϕ 'grasp'; TUK-ϕ 'poke'; TATAK-ϕ 'hit'; YUGAK-ϕ 'boil'.

[12] Probably from the original Chinese usage, some verbal nouns have both transitive and intransitive meanings like "IDOU" ('move') as in "John-ga Kuruma-o IDOU-Si-ta" (John-NOM car-ACC move(TR)-DO-PAST: 'John moved the car') and "Kuruma-ga IDOU-Si-ta" (car-NOM move(INTR)-DO-PAST: 'the car moved'). Thus, "John-ga Kuruma-o IDOU-S-ase-ta" (John-NOM car-ACC move(INTR)-DO-CAUS-PAST: 'John moved the car') has almost the same meaning as the first sentence. Nevertheless, the verbal noun itself does not demonstrate the overt morphological change. We will discuss verbal nouns further in section F.

[13] There is a prefix "Hi-" that gives a verbal noun the passive voice, such as in "Hi-Senkyo-Nin" (PASS-election(VN)-person" ('eligible person'), but it is rare and not productive.

b. Doublet Verbs[14] : TAO-s/-re 'fell/fall'; OKO-s/-r 'occur'; KIR-ϕ/-e 'cut'; MIR-ϕ/-e 'see/be.visible'; AK-e/-ϕ 'open'; TUKAM-ϕ/-ar 'hold'; SAS-ϕ/-ar 'stick'; BUTUK-e/-ar 'hit'; YUD-e/-ar 'boil'.

Table 2 is a result of a random sampling of 100 verbs from a Japanese dictionary to illustrate how doublet and non-doublet verbs distribute. This table shows the majority of Japanese native verbs (63%) have a transitive (or intransitive) doublet partner. From this table, we can see the high morphological production of transitive and intransitive verbs: 84% (=53/(53+10)) of doublet verbs have some sort of overt suffix. Although only 46% (=((53+10)/2)/((53+10)/2+37))), an estimation of the proportion of the total number of transitive) of Japanese transitive verbs have intransitive partners, virtually any Japanese intransitive verb can produce a transitive counterpart by "short causation" if not by other means of transitivization. We will discuss the meaning of this special property later in this section. As we saw, and Jacobsen (1992) wrote, "transitivity is marked by a series of verbal oppositions cutting across the native verb system" (p.56). However, this is not a peculiarity of the Japanese language alone. Many languages have similar morphological means (though not necessarily the same as those in the Japanese system) to set the transitive and intransitive apart (LINGUIST, 1996).

Table 2. Doublets and non-doublets in Japanese verbs.

V_t or V_i of Doublets	: 63	Non-Zero Ending	: 53
		Zero Ending	: 10
Non-Doublet Verbs	: 37	Transitive	: 37
Total	: 100		

[14] We adopt a notation for doublets XXX-yyy/-zzz: XXX is the root, yyy the transitive ending, and zzz the intransitive ending. "ϕ" represents a zero-morpheme.

It is notable that the difference between doublet verbs and non-doublet verbs is neither a simple phonological distinction (for counter-example, KIR-ϕ/-e 'cut' vs. KER-ϕ 'kick'; AK-e/-ϕ 'open' vs. DAK-ϕ 'hug'; MIR-ϕ/-e 'see/be.visible' vs. KOKOROMIR-ϕ 'try'), nor a superficial semantic difference (for counter-example, MIR-ϕ/-e 'see/be.visible' vs. NOZOK-ϕ 'peep at'; SAS-ϕ/-ar 'stick' vs. TUK-ϕ 'poke'; TUKAM-f/-ar 'hold' vs. NIGIR-f 'grasp'; BUTUK-e/-ar 'hit' vs. TATAK-f 'hit'; YUD-e/-ar 'boil' vs. YUGAK-f 'boil'). According to Hayatsu's (1989) study, non-doublet transitive verbs are semantically characterized as RESULT, and doublet transitive verbs as PROCESS. But, we do not pursue the semantically oriented approach in this study, and the only practical method to separate these two verb classes is to mark dictionary entries to indicate the distinction, until a computationally accountable procedure is discovered.

Now, let us move to the issue of doublet verbs. As we see in (13b), the transitivization/intransitivization suffixes in Japanese doublet verbs seem quite complex. Like the distinction between doublets verbs and non-doublet verbs, the suffix patterns are determined neither by a simple phonological difference (for counter-examples, YAK-ϕ/-e 'burn' vs. AK-e/-ϕ 'open' or YAM-e/-ϕ 'stop'; SAK-as/-ϕ 'bloom' vs. SAK-ϕ/-e 'tear'; HAG-as/-e 'peel' vs. MAG-e/-ar 'bend'; TAK-ϕ/-e 'boil' vs. TAT-e/-ϕ 'build'), nor by a superficial semantic difference (for counter-examples, OR-ϕ/-e 'snap' vs. MAG-e/-ar 'bend'; TOZ-as/-i 'close' vs. SIM-e/-ar 'close'; KOG-as/-e 'burn' vs. YAK-ϕ/-e 'burn'; SUM-as/-ϕ 'finish' vs. OW-e/-ar 'end').

The problem is whether we can give a clear definition of the functional role of each (in)transitivizing suffix attached to the root. The following presentation follows to a large degree Fujii and Kitagawa's (1997) discussion. Putting exceptional idiosyncratic patterns (e.g., NAKU-s/-nar 'loose'; KIK-ϕ/-oe 'listen/be.audible'; URUO-s/-w 'moisten'; AMAY-e/-akas 'fawn.upon/spoil'; See Jacobsen (1992)) aside, the semi-regularity of patterns is highly noticeable. From our point of view, the patterns are categorized into six classes (Table 3). This classification covers 96% of the total doublets in Jacobsen's (1992) list[15].

Table 3. Ending patterns of Japanese verbs.

({...} shows selection, and [...] means omissible.)

	Vt-end	Vi-end	Examples
A)	-[a/o]s	-{ϕ/r}	WAK-as/-ϕ (boil), UTU-s/-r (move), OYOB-os/-ϕ (influence/reach)
B)	-{a/o}s	-{e/i}	TOK-as/-e (melt), OT-os/-i (drop), TUK-as/-i (exhaust)
C)	-ϕ	-{a/o}r	SAS-ϕ/-ar (stick), TUM-ϕ/-or (stack), TUNAG-ϕ/-ar (connect)
D)	-e	"	MAG-e/-ar (bend), KOM-e/-or (push.in/hide), TOM-e/-ar (stop)
E)	-ϕ	-e	YAK-ϕ/-e (burn), NI-ϕ/-e (boil), YABUR-ϕ/-e (break)
F)	-e	-ϕ	YAM-e/-ϕ (stop), AK-e/-ϕ (open), NARAB-e/-ϕ (line.up)

To make the discussion more concrete, let us list sentences (13)-(18) as examples of the doublet types:

(13) (Type-A: -as/-ϕ)
a. John-ga Yu-o WAK-as-ta. (s-t=>sit)
John-NOM hot.water-ACC boil-TRANS-PAST
('John boiled the water.')

[15] In his list of 355 doublets (i.e., 710 verbs), 321 are regular alternations in Table 4, and 21 pairs (-s/-re, -s/-ri, -se/-ϕ, and -e/-are) can be explained by multiple formations which will be discussed later in this section. Only 13 doublets (-s/-re, -se/-ϕ, and -akas/-e) are exceptional cases. This is not an exhaustive list of the whole population in the vocabulary.

b. Yu-ga WAK-ϕ-ta. (K-t=>it)
hot.water-NOM boil-INTRANS-PAST
('The water boiled.')

(14) (type-B: -as/-e)
a. John-ga Koori-o TOK-as-ta. (s-t=>sit)
John-NOM ice-ACC melt-TRANS-PAST
('John melted the ice.')

b. Koori-ga TOK-e-ta.
ice-NOM melt-INTRANS-PAST
('The ice melted.')

(15) (type-C: -ϕ/-ar)
a. John-ga Waga-Mune-ni Naihu-o SAS-ϕ-ta. (S-t=>sit)
John-NOM own.breast-DAT knife-ACC stick-TRANS-PAST
('John plunged a knife into his own breast.')

b. Naihu-ga John-no Mune-ni SAS-ar-ta. (ar-ta=>atta)
knife-NOM John-GEN breast-DAT stick-INTRANS-PAST
('The knife plunged into John's breast.')

(16) (type-D: -e/-ar)
a. John-ga Tetu-no Boo-o MAG-e-ta.
John-NOM iron-DAT bar-ACC bend-TRANS-PAST
('John bent an iron bar.')

b. Tetu-no Boo-ga MAG-ar-ta. (ar-ta=>atta)
iron-DAT bar-NOM bend-INTRANS-PAST
('The iron bar bent.')

(17) (type-E: -ϕ/-e)
a. John-ga Shorui-o YAK-ϕ-ta. (K-t=>it)
John-NOM documents-ACC burn-TRANS-PAST
('John burned the documents.')

b. Shorui-ga YAK-e-ta. (ar-ta=>atta)
documents-NOM burn-INTRANS-PAST
('The documents burned.')

(18) (type-F: -e/-ϕ)
a. Ookesutora-ga Ensou-o YAM-e-ta.
orchestra-NOM performance-ACC stop-TRANS-PAST
('The orchestra stopped its performance.')

b. Ookesutora-no Ensoo-ga YAM-ϕ-ta. (M-ta=>nda)
orchestra-GEN performance-NOM stop-INTRANS-PAST
('The orchestra's performance stopped.')

There are four major characteristics in this categorization: i) the suffix "-as" ("-s" and "-os" are marked variants) always makes a verb transitive; ii) the suffix "-ar" ("-or" is a marked variant) always makes a verb intransitive; iii) the zero suffix (ϕ) ("-r" is a marked variant) and the suffix "-e" ("-i" is a marked variant) appear on both transitive and intransitive sides. Applying Marantz's (1984) "*No Vacuous Affixation Principle*," and defining "-as" and "-ar" as an absolute transitivizer and intransitivizer, respectively, the root categories of "-as/-ϕ"(A) and "-as/-e"(B) are categorized as intransitive, "-ϕ/-ar"(C) and "-e/-ar"(D) are categorized as transitive. Additionally, the function of zero suffix is naturally defined as "neutral," it has no effect of transitivity alternation on the root. Thus, the root of "-ϕ/-e"(E) (also correctly "-ϕ/-ar"(C)) should be classified as a transitive, and the root of "-e/-ϕ"(F) (also correctly "-as/-ϕ"(A)) must be an intransitive.

While the above accounts are so far not problematic, a serious conflict lies in the behaviors of "-e": the keenest contrast occurs between "-ϕ/-e"(E) and "-e/-ϕ"(F) - "-e" appears at both the transitive and intransitive sides, therefore it behaves as if it is both transitivizer (F) and intransitivizer (E). The notable fact is that when "-e" becomes a suffix partner of an absolute transitivizer (A) or intransitivizer (C), the root becomes a morphologically bound subordinate unit[16] which cannot stand as a syntactically independent "word." Because of this characterization, we set a two-level ordering, V̲ level (V-underbar corresponding to V^{-1} as an extension of the X-bar framework

[16] This claim seems not to hold when a root derives a "multiple," i.e., it takes more than two suffixes to alternate the transitivity. However, the mechanism of multiplication can be explained by multiple applications of doublet making, and multiples are in the minority in the doublet population.

proposed in Selkirk (1982) among others) and V (or V^0) level for transitive and intransitive separately. Therefore, roots as sub-syntactic units at (D) and (B) belong to $\underline{V}_t$ and $\underline{V}_i$, respectively (because they cannot syntactically stand by itself); also roots of (C) and (A) are categorized as V_t, and V_i, respectively (because their bare forms work as independent syntactic units). Then, how can we define the roots of (E) and (F)? Within this level ordering scheme, there are two possible solutions to this puzzling behavior of "-e." One is to assign double categories to roots of type (E) and (F), and the other is to assign a double function to the "-e" suffix. Let us take a look at each solution.

The first, the so-called "*double category*" solution, assigns both Vt and $\underline{V}$i to roots of type (E), and both Vi and $\underline{V}$t to roots of type (F). Under these root category assignments, the "-e" suffix functions to "raise" a subsyntactical $\underline{X}$ root to a syntactical X unit without changing the valency value (and its basic semantics). This functionality of "-e" can be applied uniformly, i.e., not only to types (E) and (F), but also to types (B) and (D). The difference between "-ϕ" and "-e" is that "-ϕ" cannot attach to an underbar root, but "-e" cannot attach to a non-underbar root.

The second approach, or "*double function*" model, designates only one category to every root type, but assigns two functional roles to the "-e" suffix. The basic category of the root is determined by the "null-identification" of root+ϕ, or by the "counter-identification" to the absolute assignment of (in)transitivization by "-as" or "-ar." In other words, the transitive class of a root X is calculated by either RootClass(X)=VerbClass(X+ϕ), or RootClass(X)= VTInverse(VerbClass(X+α)), where, α is "-ar" or "-as," and VTInverse(Vt)=

Vi, VTInverse(Vi)=Vt. In this model, the suffix "-e" behaves differently depending on the root's underbar status: when it attaches to a $\underline{V}$ level root, it does not change the root's transitivity nature, but removes the underbar status and surfaces it as an analytical unit. This part of the definition is exactly the same as the one in the "double category" model. However, if "-e" attaches to a non-underbar V level root, the transitivity of the derived verb is switched. That is, if the root is Vt, then the class of root+e is Vi, and if it is Vi, then root+e becomes Vt.

The root category assignments in the above two models are shown in Table 4, and the two versions of suffix functions are formalized in (19) as functions AS, AR, F, and E.

Table 4. Root category assignments in two models.

Type:	A: (-as/-ϕ)	B: (-as/-e)	C: (-ϕ/-ar)	D: (-e/-ar)	E: -ϕ/-e	F: -e/-ϕ
Root Category in Double Function	Vi	$\underline{V}$i	Vt	$\underline{V}$t	Vt	Vi
Root Category in Double Category	Vi	$\underline{V}$i	Vt	$\underline{V}$t	{Vt,$\underline{V}$i}	{Vi,$\underline{V}$t}

(19) a. AS: $X_i + \mathit{as} \rightarrow V(\mathit{as})_t$; *$X_t + \mathit{as}$

b. AR: $X_t + \mathit{ar} \rightarrow V(\mathit{ar})_i$; *$X_i + \mathit{ar}$

c. Φ: $V_n + \phi \rightarrow V_n$; *$\underline{V}_n + \phi$

d1. (double category model)
E: $\underline{V}_n + e \rightarrow V(e)_n$; *$V_n + e \rightarrow V(e)_{\tilde{n}}$

d2. (double function model)
E: $\underline{V}_n + e \rightarrow V(e)_n$; E: $V_n + e \rightarrow V(e)_{\tilde{n}}$

We now have clear definitions of derivational functions of transitivity suffixes for Japanese verbs. To solve the difference between the "double category" model and "double function" model is critical to the explanatory adequacy of linguistics. From this linguistic points of view, the arguments pro and con for each side are discussed in Fujii and Kitagawa (1997)[17]. But for our IR purpose, what is important is that both "double category" and "double function" versions of transitivity alternation models are descriptively and computationally equivalent. The only difference is that the "double category" system has slightly more the flavor of a data-driven mechanism, and the "double function" system has a stronger rule-driven mechanism. So, at the implementational level, the "double function" system puts less load on the dictionary because of its additional operational cost. However, the extra load on the dictionary in the "double category" system will be not extremely high, because the number of stored roots in the dictionary (lexicon) is not affected, and the only added information is the transitivity markers for type (E) (i.e., -ϕ/-e) and (F) (i.e., -e/-ϕ) roots. Nevertheless, the basic perspective of either model is represented and maintained by a set of rules, and this fundamental property makes the system framework completely different from a system which relies completely on the permanent lexicon such as Juman (Matsumoto, et al., 1991). So, putting overly detailed linguistic arguments aside, let us adopt the "double-function" mechanism as a working "implementational short-cut."

Now, let us ask what the significance is of these morphological

[17] In this paper, authors went through the argument touching on various points, such as historical perspective, cognitive efficiency, cross-linguistic contrastive examination, unaccusative representations, and formation of multiples. We will touch on unaccusative issues later in this section. Concerning multiples such as a triplet YAM-e/-{ϕ,ar}, we treat them simply as combinations of doublets - in this example, YAM-e/-ϕ and YAM-e/-ar.

classifications for the application of information retrieval. To answer this question, we have to address the "interface problem," i.e., justification of the boundary of two linguistic components, in our case, between word formations for lexicon and syntactical operations on word arrangement. We consider this to be so because even from the most modest view of IR, lexical and syntactic effects on retrieval performance are said to distinctly differ from each other.

We already know that there are basically three types of suffixed (including -ϕ) transitive verbs, that is, X+ϕ, X+e, and X+as. A question is whether we may treat them (theoretically or practically) in a homogeneous manner as a single transitive kind, or not. Also, can we treat the three types of intransitive verbs, X+ϕ, X+e, and X+ar (theoretically or practically) in the same manner? First, let us consider the case of transitivizers. In contrast to the traditional views of Japanese transitivity alternations (e.g., they are "unproductive" (Jacobsen, 1992), and "underived" (Baker, 1988)), our framework for doublets demonstrates the opposite, i.e., "productive" and "derivational." Furthermore, we easily recognize the similarity between the transitive suffix "-as" and causative morpheme "-ase," as well as between the intransitive suffix "-ar" and the passive morpheme "-are." Additionally, some connection between the transitivity suffix "-e" and the potential morpheme "-[rar]e" may be pointed out. Fujii and Kitagawa (1997) conceptualized these parallel relationships as "short causative/passive/ potential" versus "long causative/passive/potential." The productivity of the short causative is especially prominent. This is a phenomenon which occurs across the boundaries of lexicon and syntax. As Shibatani (1973, 1976) discussed, a short causative occasionally demonstrates lexical characteristics (e.g., interpretation of a phrase with a reflexive), but if a transitive entity already exists in a permanent lexicon (which "blocks" short causatives from entering into a *paradigmatic structure* (PDS) in Miyagawa's (1989) framework), the short causative has the same semantic transparency which the long causative of syntax has, so there is no idiosyncrasy. Thus, in the latter case, short causatives and long causatives are in the relationship of allomorphs, and they are stylistic variants in a sentence. Since we now have a systematic derivational framework of transitivity alternation, we can restate this assertion: If a short causative is not blocked at a transitive slot of PDS, the intransitive slot must be occupied by a form with "-ϕ" or "-e" as a type of (A) or (B), respectively.[18] (It is unlikely to be empty there because short causativization is a highly marked form, and an unmarked or less marked form is expected to exist.) Furthermore, when a short causative is lexical, and it is allowed to stay at the transitive slot of PDS, the root must be intransitive (i.e., Vi or $\underline{\text{V}}$i) in either the double-function or the double-category model, or in other words contrapositionally, if a root is transitive (i.e., Vt or $\underline{\text{V}}$t), the short causative is syntactical by being blocked from entering into the permanent lexicon. Note that the converse does not hold: even if the root is Vi, its short causative formation may be still syntactical because it can be blocked when the suffix pattern is -e/-ϕ. (In the double-category system, it is not necessary to consider the attachment of a short causative to Vt because of Marantz's (1984) "*No Vacuous Affixation Principle*.") Therefore, the transitivity category of a verbal root that is a property permanently recorded in the lexicon, is not the sole factor in determining whether a short causative

[18] We are not considering irregular ending patterns in this discussion for the sake of simplicity.

is lexical. Some other morphologically productive lexicalizations must also be coordinated to make the (in)transitivity assignment definitive.

The next question is whether we can construct similar arguments for "short passive" and "short potential." Although we do not so far have sufficient evidence in the linguistic literature, the answer to this question seems to be negative, at least to some degree. In (20), a type (F) verb NARAB-e/-ϕ 'line.up' shows a transitive case of paradigm blocking. Thus, the transitive form NARAB-e is lexicalized, and has an idiosyncratic meaning 'to express complaints' in (20b,c). Hence, the short and long causatives (20d) are blocked from entering into the permanent lexicon, and equally receive the causative sense of the standard meaning transparently from their unmarked intransitive form (20a).

(20) a. Heitai-ga/*Huhei-ga NARAB-ϕ-ta. ('B-t=>nd')
soldier/complaint-NOM line.up-INTR-PAST
('The soldiers lined up.')

b. *Taichou-ga Heitai-o NARAB-e-ta.
commander-NOM soldier-ACC line.up-TR-PAST
('The commander lined up the soldiers.')

c. Taichou-ga Huhei-o NARAB-e-ta.
commander-NOM complaint-ACC line.up-TR-PAST
('The commander complained a lot.')

d. Taichou-ga Heitai-o/*Huhei-o NARAB{-as, -ase}-ta. (s-t=>sit)
commander-NOM soldier/complaint-ACC line.up-CAUS-PAST
('The commander let the soldiers/complaints line up.')

On the other hand, (21) shows an intransitive case. A doublet YABUR-ϕ/-e 'break' is type (E), and the transitive form YABUR-ϕ receives an idiosyncratic meaning of 'violate (rule, law)' (21a: "Kisoku-o" [rule-ACC]) as a lexicalized item. The transitive form YABUR-e cannot receive the idiosyncratic sense (21c), but only the standard meaning (21a: "Kago-o," and 21b). Therefore, similarly to the causative case, forms with a short/long passive suffix (21d) are blocked from coming into PDS, and both standard and idiosyncratic meanings should be conveyed transparently as syntactically passivized. However, the similarity ends here. Unlike the causative situation where both short and long causatives were grammatical, the short passive is not grammatical (21d: *YABUR-ar), although the long passive is (21d: YABUR-are), as expected.

(21) a. Heitai-ga Kago-o/Kisoku-o YABUR-ϕ-ta. ('R-t=>tt')
soldier-NOM basket/rule-ACC break-TR-PAST
('The soldiers broke the basket/rule.')

b. Kago-ga YABUR-e-ta.
basket-NOM break-INTR-PAST
('The basket broke.')

c. *Kisoku-ga YABUR-e-ta.
rule-NOM break-INTR-PAST
('The rule broke.')

d. Kago-ga/Kisoku-ga (Heitai-niyotte) YABUR{*-ar, -are}-ta.
basket/rule-NOM (soldier-Caused.by) break-PASS-PAST
('The basket/rule was broken (by the soldier).')

What is the situation of potentials? In this case, the mechanism of blocking does not involve the realization of potential verbs in any way. What affects it most is that none of the other suffixes, i.e., "-as" or "-ar," can substitute "-e" as an alternative short potential suffix. This is a quite different situation from where "-as" and "-ar" had "-e" as a possible alternative in their roles as transitivizer and intransitivizer, respectively. Furthermore, semantically speaking, the potential (including its related meanings such as "zihatu" 'spontaneous occurrence') is an additional property of either transitives or intransitives, and no verb exists without (in)transitivity. Thus, if the doublet is type (C), e.g., SAS-ϕ/-ar which has V_t root category, attaching an intransitive "-e" to the root is not only morphologically ungrammatical[19] because the intransitivity slot in PDS is already occupied by SAS-ar, but also it makes the identical surface form to the potential of the transitive, i.e., SAS-ϕ.e 'can stick.' Similarly, if the type is (A), e.g., HAGEM-as/-ϕ, replacing "-as" with "-e" does not make another bivalent form, but produces a form the same as the potential intransitive form HAGEM-ϕ-e 'can be diligent.' When a short potential occupies the transitive or intransitive slot, and its partner has a zero suffix, i.e., the cases of type (B) and (D), similar ambiguity occurs such that NARAB+e can be interpreted as either a transitive ('line up (something)' as NARAB-e), or potential intransitive ('can line up' as NARAB-f-e). Thus, YABUR+e can have either an intransitive meaning ('break' by YABUR-e), or a potential transitive meaning ('can break (something)' by YABUR-f-e). In any case, the suffixation of "-e" introduces a *valency ambiguity* instead of a *valency alternation*.

(22) consists of schematized figures of the overall discussion of blocking relationships. From the point of view of the interface problem, each of the three suffixes, i.e., short causative, short passive, and short potential, has unique characteristics. Regarding the short causative, no matter whether it is lexicalized or works syntactically, it is grammatical and functions to increase valency. This is not like the short passive, which becomes ungrammatical

[19] Although there is a possibility to produce a multiple, having a suffix pattern -ϕ/-{ar,e} is unlikely compared to the standard -e/-{ϕ,ar} (e.g., YAM-e/-{ϕ,ar} 'stop') or -{ϕ,as}/-e (e.g., YUR-{-ϕ,as}/-e 'shake') as described in Fujii & Kitagawa (1997). Indeed, we could not find any example of -ϕ/-{ar,e} pattern.

when it is blocked, nor like the short potential, which causes a valency ambiguity. Concerning the short passive, it decreases valency when it is lexicalized, but cannot exist as a syntactical element. The short potential, unless it is lexicalized and functions as an (in)transitivizer coupled with short causative or passive, causes valency ambiguity[20] .

(22)

a. Blocked Short Causative

```
          Vt                     Vi
   +------------------------+--------------------------+
        root+e                 root+ϕ                  root=type(F)={NARAB, AK}
   +------------------------+--------------------------+
  blocked:{-as. -ase}
```

b. Unblocked Short Causative

```
   +------------------------+--------------------------+
        root+as                root+{ϕ,e}              root=type(A,B)={HAGEM, TOK}
   +------------------------+--------------------------+
```

c. Blocked Short Passive

```
   +------------------------+--------------------------+
        root+ϕ                 root+e                  root=type(E)={YABUR, NI}
   +------------------------+--------------------------+
                          blocked:{*-ar. -are}
```

[20] In the end, we have four different kinds of "potentials," as seen in the following examples:
1) (Lexical) short potential: Ie-ga YAK-e-ru. [house-NOM burn-INTRANS-PRES]
2) Syntactical short potential: Hari-o SAS-ϕ-e-ru. [needle-ACC stick-TRANS-POTN-PRES]
3) Syntactical long potential: Eda-o MAG-e-rare-ru. [branch-ACC bend-TRANS-POTN-PRES]
4) Compounding potential: Sensou-o OK-os[i]-e-ru. [war -ACC start-TRANS-POTN-PRES]
Although we denote the lexical "-e" in 1) as simply a "short potential" in corresponding to the "short causative" for the sake of terminological consistency, it is actually a "spontaneous" intransitive as discussed here. Syntactical 2) and 3) are often exchangeable as stylistic variants in discourse such as in:
5) John-ha Musiba-de Katai Mono-ga TABE-{re/rare}-naku-nar-ta. (r-t=>tt)
('John cannot eat hard food because of his cavities.')
The fourth pattern of potential is probably "a loan expression from Chinese, ... and is used primarily in a literary context with KANGO (Chinese compound) flavor." (Kitagawa, personal communication, 1997) So, this pattern is not common in spoken language, but a formal text like the newspaper article in our experiments might display it with enough frequency to have some effect on retrieval. We will take these patterns into account in the experimental query formulation which will be discussed in the next chapter.

d. Unblocked Short Passive

```
+------------------------+---------------------------+
     root+{ϕ,e}               root+ar              root=type(C,D)={SAS, MAG}
+------------------------+---------------------------+
```

e. Blocked Short Potential

```
+------------------------+---------------------------+
     root+ϕ                  root+ar
     root+as                 root+ϕ                root=type(A,C)={SAS, HAGEM}
+------------------------+---------------------------+
  "root+e" results in a valency ambiguity with potential.
```

f. Unblocked Short Potential

```
+------------------------+---------------------------+
     root+ϕ                  root+e
     root+e                  root+ϕ                  root=type(E,F)={YABUR,NARAB}
+------------------------+---------------------------+
 "root+e" results in a valency ambiguity with potential.
```

Although the number of suffixes for transitivity alternations is basically only four, i.e., "-ϕ," "-as," "-ar," and "-e," we realize the complexity of the system once we look at their lexicality and syntacticality, idiosyncrasy and productivity, etc. Their formal properties are associated with certain semantic characterizations in discourse, and consequently will affect the retrieval through not completely knowable cause-and-effect chains. Simply specifying the morphological root and the indicator of transitivity (or the suffix itself) is not enough to capture such significant complexities of the Japanese verbal system.

Although the issue we have discussed in this section appears far from detailed from the point of practicality, it is very important for grasping possible analyzable subdomains, because such knowledge may help our performance analysis by suggesting a possible underlying mechanism of effecting on the retrieval process, as well as indicating future opportunities to solve further retrieval problems. And as already noted, Japanese is not the only language to show a complicated verbal morphology for valency changes. Many languages have an overt and systematic morphological system for transitivization/causativization and intransitivization/passivization. (LINGUIST, 1996) Korean, for example, demonstrates a similar double function of alternating transitivization and intransitivization by a single morpheme. (Fujii and Kitagawa, 1997)

E. Syntactical Valency Control with Passivization and Causativization

In the previous section, an issue related to syntax was discussed when the "interface" problem between lexicon and syntax was discussed. In following discussion, we will approach the same problem from a different direction - how easy or difficult it is to paraphrase a sentence into a different construction. Although paraphrasing is a task of semantic manipulation, we will try to find an approach that is formally and structurally accountable. In generative grammar, the *Projection Principle* considers that certain aspects of he meaning of the sentence are full-fledged as interpretable in the underlying structure (D-structure) even before the surface structure (S-structure) is derived from the D-structure. So, let us examine the verbal construction of sentences in terms of D-structure.

In the "*Unaccusative Hypothesis*" (Perlmutter, 1978, Levin and Rappaport, 1995), intransitive verbs are further classified into two distinct groups, called *unergative* and *unaccusative* verbs. Thus, a verb must be either unergative, unaccusative, or transitive (which includes ditransitive). The external argument of an unergative verb is placed at the subject position both in D-structure and S-structure, and the object position remains empty at both levels. On the other hand, in this hypothesis, an unaccusative verb does not have an external argument, and the internal argument of an unaccusative verb is thought to occupy the object position in the D-structure. Typical theta-role arrangements in argument structures of these three verb types are shown in (23).

(23) a. unergative: (Agent < ϕ >)

b. unaccusative: (ϕ <Theme>)

c. transitive: (Agent <Theme>)

We already pointed out two parallelisms between intransitive and passive, and between transitive and causative based on their valencies so that we can implement valency control paraphrasing as we saw in (1), (2), and others. These relationships hold no matter the intransitive verb is either unergative or unaccusative. Furthermore, Levin and Rappaport (1995) noted about the relationship between passivization and unaccusativity: "on this definition [i.e., taking an internal argument but no external argument, discussed in Perlmutter (1978)], unaccusative verbs are identical in D-Structure configurational terms to passive verbs" (p.3). Consequently, both sentence patterns consistently maintain their Theme (or Patient) roles because of their homologous structures.

However, the situation is quite different in the unergative causative case versus transitive. As Shibatabi (1973) argued, in an unergative causative sentence, an unergative sentence is embedded in a matrix clause, therefore the causee can maintain its agentivity which cannot be found in the transitive counterpart. Thus, both the patient marker (-o) and the agentive marker (-ni) can be attached to the object of the causative verb (24a. and b.)

though the transitive (i.e., short causative of type A or B) sentence allows only the patient marker but not the agentive one (24c. and d.). In other words, using this causative relationship for paraphrasing utilizes more distance connection in both syntax and semantics than passivization.

(24) a. Kanojo-wa Kodomo-o asahayaku OK-i-sase-ta.
she-TOP child-acc(PAT) eraly.morning wake.up-INTR-CAUS-PAST
('She made her child wake up in the early morning.')

b. Kanojo-wa Kodomo-ni asahayaku OK-i-sase-ta.
she-TOP child-dat(AGENT) eraly.morning wake.up-INTR-CAUS-PAST
('She let her child wake up in the early morning.')

c. Kanojo-wa Kodomo-o asahayaku OK-os-ta. (s-t>sit)
she-TOP child-acc(PAT) eraly.morning wake.up-TR-PAST
('She woke her child up in the early morning.')

d. *Kanojo-wa Kodomo-ni asahayaku OK-os-ta. (s-t>sit)
she-TOP child-dat(AGENT) eraly.morning wake.up-TR-PAST
('She woke her child up in the early morning.')

Next, let us look at the distribution of these verb types in doublet patterns. We already saw that the majority of Japanese verbs have a doublet partner, and the productivity of the short causative eventually leads to any intransitive verb having a transitive partner. (25) is a list of unergative verbs classified by doublet patterns. In this list, we first notice that unergative verbs do not distribute uniformly over doublet patterns, and especially that there is a concentration in type (A) (and type(B) as its variant), where the short-causative morpheme "-as" is attached. This is, to some degree, reasonable because one characteristic of unergative verbs is a demonstration of the agent's will or volitional state, and short-causation can naturally add the meaning of an external force operating against the will. Even though usually no other alternative ending is attached to make another transitive form for type (A) and (B) so that a short-causative behaves as a lexical item, nevertheless some causative sense (if not, at least a cognitive association to the long (i.e., syntactical) causation because of morphological or phonological relatedness) may be indicated in language performance.

(25) Examples of Unergative Verbs

a. (type-A [Vi: -as/-ϕ]):
NAK-ϕ 'cry', SAWAG-ϕ 'make noise', NE-ϕ (-kas/) 'sleep',
UGOK-ϕ 'move', WARAW-ϕ 'laugh', HOHOEM-ϕ 'smile',
NEMUR-ϕ 'fall into sleep', SEKIKOM-ϕ 'cough', KAYOW-ϕ 'commute',
ISAM-ϕ 'be brave', ABARE-ϕ 'be violent', DAMAR-ϕ 'be quiet',
ODOR-ϕ 'dance', SUWAR-ϕ 'sit', HANE-ϕ 'jump', TOB-ϕ 'fly',
ARUK-ϕ 'walk', AYUM-ϕ 'walk', HASIR-ϕ 'run', HAW-ϕ, 'crawl',
YOR-ϕ 'approach', OYOG-ϕ 'swim', KOROB-ϕ, 'tumble down',
HATARAK-ϕ, 'work', ASOB-ϕ 'play', SINOB-ϕ 'hide',
SYABER-ϕ 'speak', INOR-ϕ 'pray', HOE-ϕ 'bark', OKOR-ϕ 'get angry',
SIN-ϕ[21] , 'die', TOTUG-ϕ 'get married', HUR-ϕ[22] 'fall (rain, snow, etc.)',
SUM[AW]-ϕ 'reside', AW-ϕ 'meet', IK-ϕ 'go', YUK-ϕ 'go'

b. (type-B [Vi: -as/-e]):
NIG-e 'escape', OR-i (-os/) 'go down', D-e 'exit', OK-i (-os/) 'wake up',
BAK-e 'disguise'

c. (type-C [Vt: -ϕ/-ar]):
KURUM-ar 'cover (wear)'

d. (type-D [Vt: -e/-ar]):
MAG-ar 'bend (turn a curve/corner)', TOM-ar 'stop (stop walking/running)',
KUWAW-ar 'be added (join in a group)', YOKOTAW-ar 'lie (the own body)',
MAZIW-ar 'be mixed (get acquainted)', TUK-ar 'soak (take a bath)'

e. (type-E [Vt: -ϕ/-e]):
?KUDAK-e 'break (show a relaxed attitude)'

f. (type-F [Vi: -e/-ϕ]):
KAGAM-ϕ 'stoop', UTUMUK-ϕ 'hang down own head',
YASUM-ϕ 'rest', SUSUM-ϕ 'go forward', DOK-ϕ 'step aside',
[HA]IR-ϕ 'enter', NARAB-ϕ 'line up', TAT-ϕ 'stand.up'

As we see in (25), the existence of a short-causative morpheme does not necessarily have to be tied to the condition of unergativity. We can find unergative verbs in other doublet patterns. This is true especially when they received idiosyncratic lexical meanings. Unergative verbs with straight-

[21] This classification follows Kageyama's (1995) analysis. He pointed out that verbs which have the superficially identical meaning 'die' should be categorized differently: SIN-ϕ as unergative, NAKUN-ar as unaccusative.

[22] This classification follows Kageyama's (1995) analysis.

forward use (i.e., no idiosyncratic deviation) are hardly found in types (C, D, E) as in (25c, d, e), and they tend to have some voluntary senses (e.g., MAG-ar 'turn (one's own body)'[23] , KUWAW-ar 'join (in a group/party),' etc.) which are semantically or analogically derived from a more general involuntary, natural, or neutral meaning, or a more or less idiosyncratic meaning such as 'bend' and 'add,' respectively. This situation may be understood by their roots being categorized as transitives, i.e., Vt or $\underline{V}$t. On the other hand, unergatives of Type (F) seem to behave more similarly to type (A), probably because both roots are Vi. They describe kinds of bodily processes (e.g., *cry, laugh, move, jump,* etc. for type (A); *stoop, hang down own head, stand up,* etc. for Type (F)) as their unique meaning, so they somehow differ from types (C, D, E). However, there is one common characteristic between type (D) and (F): their transitive partner (root+"e") often fails to keep a "self-centered" sense, such as MAG-e 'bend (branch, board, etc.),' KUWA-e 'add (thing, number, etc.),' TAT-e 'build (house, etc.),' or NARAB-e 'arrange (things)' - so if we want to keep the semantic transparency coming from the intransitive sense, we should instead use causative forms (by syntax), such as MAG-ar-as[e] 'have turn,' TAT-as[e] 'have stand,' or NARAB-as[e] 'have people line up,' respectively, rather than their transitive counterparts. Figure 1 indicates a conceptual model of verb type distribution.

[23] We can confirm the difference between the unergativity and unaccusativity by Miyagawa's (1989) test which uses an adverbial insertion accompanied with a numeral quantifier such as:
(26) a. Kuruma-ga (*Kousaten-o) 3-dai MAG-ar-ta. ('Three cars turn at the crossing loudly.')
car-NOM (crossing-LOC) three turn-INTR-PAST
b. Tetunoboo-ga (Otootatee) 3-bon MAG-ar-ta. ('Three iron rods bent loudly.')
iron.rod-NOM (loudly) three bend-INTR-PAST
Note that MAG-ar in (25a) is still an intransitive verb because not capable of being passivized by raising the object (as we intend to use this characteristic for paraphrasing), although it takes Kousaten 'crossing' as an obligatory locational argument. In other words, the argument structure of MAG-ar in (26a) is (AGENT <LOC>) in contrast to (26b)'s MAG-ar: (ϕ <THEME>) as an unaccusative.

The distinction between unaccusativity and unergativity for intransitive verbs affects the ease of making a paraphrasal causative or passive sentence. Let us consider these two constructions in separate discussions.

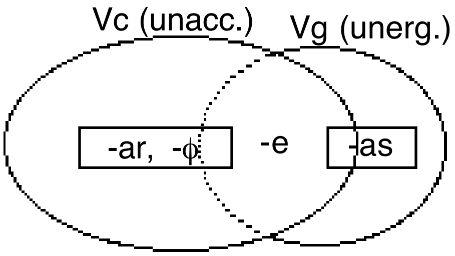


Figure 1. A conceptual model of verb type distribution.

1. Causativization

In causativization, the domain of the paraphrase operation must be intransitive, in other words, either unaccusative or unergative, to produce an equivalent sentence with a transitive verb. It is important to distinguish the retrieval effects of the long ("-ase") and short ("-as") causatives if their difference is not of the stylistic variants (i.e., when PDS's transitive slot is occupied by zero or "-e"), but of the linguistic modules (i.e., syntactical and lexical, respectively). For an unergative verb, typically a type (A) verb, causativization of either the short or long sort is straightforwardly acceptable because it simply adds a causer ('Emily' in (27a)) to the original agent ('John') which has some voluntary state ('cry'), though they show some semantic difference as seen in (27b,c). Shibatani (1973) proved that short and long causatives, if they are not blocked, have lexical and syntactical origins, respectively.

(27) a. John-ga Emily-o NAK-{as, ase}-ta. (s-t=>sit)
Emily-NOM John-ACC cry-CAUS-PAST
('John made/let Emily cry.')

b. John-ga Inu-no Sippo-o Hunde NAK-{as, ?ase}-ta. (s-t=>sit)

John-NOM dog-GEN tail-ACC step-Adv cry-CAUS-PAST
('John made the dog cry by stepping on its tail.')

c. John-ga Inu-o Ie-no Soto-de NAK-{?as, ase}-ta. (s-t=>sit)
John-NOM dog-ACC house-GEN outside-LOC cry-CAUS-PAST
('John let the dog cry outside the house.')

What is the case of unaccusative verbs? Although there are various semantic groups in unaccusative verbs, as Perlmutter and Postal (1984) listed, and so it is not clear whether a definitive characterization for this purpose is possible, one distinctive characteristic of the typical unaccusative verb is that the subject of the sentence has a PATIENT role that should be already expressed in the semantics of the root as the core of the verbal concept. Consequently, the one straightforward and immediate way to realize a bivalent relationship in a sentence is to derive a transitive form by a transitivizer rather than by syntactical causativization. So, if transitivization does not use the short causative morpheme ('-as'), in other words '-e' or '-ϕ' is used for Type (C,D,E,F), the (long) causativized construction which shares the semantic characteristics, to some degree, with short causative formation, will show some markedness, if it is not in fact ungrammatical. To substantiate this prediction, let us look at the application of each doublet pattern in the following paragraphs.

(28) is an example of type (E) (i.e., -ϕ/-e pattern) with a verb OR-ϕ/e 'bend'. A plain causative sentence (28a) is hard to accept even though the standard transitive case (28b) is grammatical. However, it is relatively easy to make a causative sentence by adding some external or indirect context such as in (28c). Thus, our prediction is right in this pattern. Similar type (E) verbs are KIR-e 'cut', SAK-e 'tear', NEZIR-e 'twist', YAK-e 'burn', MUK-e 'peel', TOK-e 'solve', KUDAK-e 'break into pieces', MI-e 'be visible', among others.

(28) a. ?*John-ga Eda-o OR-e-{sas, sase}-ta.
John-NOM branch-ACC break-INTR-CAUS-PAST
('John made the branch break.')

b. John-ga Eda-o OR-ϕ-ta. (r-t=>tt)
John-NOM branch-ACC break-TRANS-PAST
('John broke the branch.')

c. John-ga Edasaki-ni Omori-o Tukete Sono-Eda-o OR-e-{?sas, sase}-ta.
John-NOM End-LOC weight-ACC DET-branch-ACC break-INTR-CAUS-PAST
('John made the branch break by putting a weight at the end.') (29) is an example of type (F) (i.e., -e/-ϕ) bivalent sentences with a verb SIZUM-e/-ϕ 'sink'. (Similar type (F) verbs are YUGAM-ϕ 'twist', YURUM-ϕ 'loosen', TAWAM-ϕ 'bend', UKAB-ϕ 'float', KATAMUK-ϕ 'tilt', URAMUK-ϕ 'turn up-side-down', TAT-ϕ[24] 'stand', SAKADAT-ϕ 'stand against a direction', among others.) Here, the sense of indirectness of the causative action in (29a) is less noticeable than in (28a). Although (29a) still gives an impression of external coercive force in the event as oppose to a natural process, a plain causative sentence of this type with no helping adverbial phrase as in (29a) is, in author's judgment, perfectly acceptable, with a spontaneous[25] flavor in contrast to the hardly acceptable case of (28a).

(29) a. Otokonoko-ga Omocha-no Booto-o SIZUM-ϕ-{as, ase}-ta.
boy-NOM toy-GEN boat-ACC sink-INTR-CAUS-PAST
('A boy made a toy boat sink.')

b. Otokonoko-ga Omocha-no Booto-o SIZUM-e-ta.
boy-NOM toy-GEN boat-ACC sink-TRANS-PAST
('A boy sank a toy boat.')

The meaning gap between (29a) and (29b) (and also between (28b) and (28c)) has a resonance with McCawley's (1978) analysis of speech acts: the

[24] The sentence "*John-ga Ie-o TAT-as[e]-ta" ('John made a house built.') is not acceptable because this verb TAT-ϕ signifies a sort of uncontrollable emergence rather than ordinary actions or states. When the verb has a meaning of 'stand' rather than 'build,' it becomes acceptable like "Koronbusu-ga Tamago-o TAT-as[e]-ta" ('Columbus made an egg stand.') The verb AK-ϕ has similar characteristics. In this case, AK-as has an idiosyncratic meaning of 'reveal,' which blocks the use of the causative meaning.

[25] Consequently, the semantics of causative versions tend to indicate the imperfect or durative aspect, but the transitive versions tend to show the resultative aspect. For example, we can say "Kare-wa Booto-o 1-punkan SIZUM-ase-ta" ('He let a boat sink for one minute.'), but it is difficult to accept "?*Kare-wa Booto-o 1-punkan SIZUM-e-ta" ('He has sunk a boat for one minute.').

shorter form is used for more direct meaning, the longer for indirect description, as a sort of "division of labor." However, this does not explain the unacceptability of (26a). The above cases seem to share the same nature with triplet formations like KARAM-{e,as[e]}/-ϕ which was reported by Fujii and Kitagawa (1997). The reason for this acceptability of (29a), in contrast to (28a), might have some connection to the characteristics of the unergative category (i.e., type (F) is similar to (A), but differs from (E). See (25) and its observations.) as we described before.

The next question pertains to whether type (C) (i.e., -ϕ/-ar) or (D) (i.e., -e/-ar) is similar to the previous situation of type (E) (e.g., (26)), or to that of type (F) (e.g., (29)) (See Table 5). (30) and (31) are examples of type (C) and (D), respectively. Again, without the help of an adverbial phrase, they resist being accepted, giving a sense of lack of an obligatory element. Thus, these cases resemble type (E). The same kind of type (C) verbs are HASAM-ar 'nip', HUSAG-ar 'clog', KURUM-ar 'wrap', SAS-ar 'stick', and type (D) are HAM-ar 'fit', HIROG-ar 'spread', KIM-ar 'be decided', OSAM-ar 'be settled', TOM-ar 'stop' (involuntary), SOM-ar 'color,' TASUK-ar 'be helped.'

Table 5. Root types and the causative acceptability
of intransitive verbs.

Verb Types	Acceptability of Intransitive Causative
C, D, E	**X** Resist being causativized
F	Δ Acceptable, but differs from the transitive meaning
A,B	**O** Acceptable, and have close meaning to transitive

(30) a. ?*John-ga 2honno Paipu-o TUNAG-ar-{as, ase}-ta.
John-NOM two pipe-ACC connect-INTR-CAUS-PAST
('John made the two pipes connect.')

b. John-ga 2honno Paipu-o TUNAG-ϕ-ta. (G-t>id)
John-NOM two pipe-ACC connect-TRANS-PAST
('John broke the branch.')

c. John-ga Teepu-de 2honno Paipu-o TUNAG-ar-{?as, ase}-ta.
John-NOM tape-INSTR two pipe-ACC connect-INTR-CAUS-PAST
('John made the two pipes connect using the tape.')

(31) a. ?*John-ga Eda-o MAG-ar-{as, ase}-ta.
John-NOM branch-ACC bend-INTR-CAUS-PAST
('John made the branch break.')

a'. John-ga Untenshu-ni sono Kado-o MAG-ar-{as, ase}-ta.
John-NOM driver-DAT the corner-LOC turn-INTR-CAUS-PAST
('John made the driver turn the corner.')

b. John-ga Eda-o MAG-e-ta.
John-NOM branch-ACC bend-TRANS-PAST
('John broke the branch.')

c. John-ga Edasaki-ni Omori-o Tuke-te Eda-o MAG-ar-{?as, ase}-ta.
John-NOM End-LOC weight-ACC put-ADV branch-ACC break-INTR-CAUS-PAST
('John made the branch break by putting a weight at the end.')

What we observed is that a syntactic causative (because the transitive already exists as another form) of a transitive root[26] (type (C, D, E), e.g., (30, 31, 28)) is unacceptable with a neutral interpretation, but for an intransitive root (type (F) e.g., (29)), the same construction becomes acceptable. We may consider that this situation is caused by some phonological or morphological constraints as Kageyama (1993) once suggested the unacceptability of an intransitivizer "-e" plus passivizer "-rare" for that reason. But, in our case, the simple phonological condition of a concatenation of two morphemes, "-e" and "-as[e]" may be not sufficient because (28c) became acceptable. Furthermore, the former analysis seems to accord with the unaccusative case in type (A) and (B) as we see in (32) and (33), respectively. In this case, the

[26] If we consider double categories as encoded in the root of type (E) as Vt and V̲i (also type (F) as Vi and V̲t), we can regard the non-underbar status as the principal category.

long causative, which is a syntactic construction, is ungrammatical in a plain sentence (32a, 33a), but again with an adverbial phrase added for spontaneous meaning, the sentence becomes acceptable as in (32b, 33b). Furthermore, this relationship holds even when the form is morphologically marked, such as OT-os and OT-i-sase in (33), so, it stands to reason it is stored as a single unit in the permanent lexicon, and the phonological and morphological similarity (not like that between "-as" and "-ase") cannot be given credit for the effect.

(32) a. Emily-ga Sentaku-o KAWAK-{as, *ase}-ta. (s-t=>sit)
Emily-NOM laundry-ACC dry-CAUS-PAST
('Emily made/let the laundry dry.')

b. Emily-ga Kabe-no Penki-o Sizenni KAWAK-{?as, ase}-ta. (s-t=>sit)
Emily-NOM wall-GEN paint-ACC naturally dry-CAUS-PAST
('Emily let the wall paint dry in a natural way.')

(33) a. John-ga booru-o Biru-kara OT-{os, *i-sase}-ta. (s-t=>sit)
John-NOM laundry-ACC fall-{TRANS, INTR-CAUS}-PAST
('John dropped a ball from the building.')

b. John-ga Kabe-no Doro-o Hi-ni Atete Sizenni OT-{?os, i-sase}-ta. (s-t=>sit)
John-NOM wall-GEN mud-ACC sunlight-DAT apply-Adv naturally
fall-{TRANS, INTR-CAUS}-PAST
('John let the mud on the wall fall in a natural way by putting it under
the sunlight.')

To summarize about the causativization, we observed that when the root is transitive for causativization of an unaccusative verb (type (C, D, E)), the acceptability gap is a matter of a dichotomy of lexical versus syntactical. If the type is (A) or (B), this lexical vs. syntactical situation is translated as a matter of short causative versus long causative. For both unaccusative verbs with intransitive roots (type (F)), and most unergative verbs, the distinction of short and long causatives has weak impact on acceptability, therefore predictably, also on the retrieval performance.

2. Passivization

We should first note that a large part of passivization creates no problem for paraphrasal operation because the domain of the operation is transitive verbs, and the essential operation is simply to raise the item at the object position to the subject position. As we saw in (23), transitive verbs have a unique argument structure in contrast to two possible structures for intransitive verbs, i.e., one for the unaccusative class, the other for the unergative class. In other words, passivization for transitive verbs is a uniformly applicable operation. Although, in Japanese, even a sentence with an intransitive verb can be passivized, which is the so-called "indirect passive"[27] (the ordinary passivization for transitive verbs is called "direct passive"), this simply does not affect our paraphrasal operation because of our transitive domain for valency paraphrasing. So, for our purposes we put this problem aside.

One possible problem is that transitive verbs of type (A) (i.e., -as/-ϕ) and (B) (i.e., -as/-e) have a short causative morpheme, so the passivization of these verbs should be addressed as a part of the causative-passive situation. Thus, causativization and passivization might interact with each other. (By the same token, we may see a parallel in cases of intransitive verbs with short passive morpheme (i.e., passive-causative) with type (C) and (D) as we already analyzed.)

[27] According to Kageyama (1995), an unaccusative sentence cannot be passivized even in an indirect way. However, it is possible to make acceptable examples, such as: "Ie-no Mae-no Ki-ni OR-e-rare-te, Heya-ni Hi-ga Sasuyouni-natta" ('Being broken (intr.) by the tree in front of the house, the sun light now comes in my room.'); "Sono Kikai-ni KOWA-re-rare-te, Kouzyou-no Seisan-wa Zenbu Tomatta" ('Being broken (intr.) by the machine, all the factory production has stopped.'). Therefore, theoretically speaking, we can paraphrase these sentences with unaccusative verbs by passivization. However, since such construction does not occur frequently (especially in formal texts like newspaper articles in our experiments), we will put it aside from our consideration.

In Kageyama's (1993) study, causative-passive construction was analyzed associated with the dichotomy of unaccusative and unergative. According to his analysis, a sentence with an unergative verb can be causative-passivized, but one with an unaccusative verb cannot at least in a direct way of passivization.

In unergative cases, his prediction seem to be correct, though there is occasional preferable selection of either a long or short causative form as shown in (34a,b).

(34) a. John-wa Kinzyo-no Gakidaishou-ni NAK-{as, ?ase}-[r]are-ta.
John-NOM neighbor-GEN bully-AGENT cry-CAUS-PASS-PAST
('John was made to cry by a bully of a neighbor.')

b. John-wa Emily-ni Ookina-Heya-ni Hitori-Hootteokarete
Kinosumu-made NAK-{?as, ase}-[r]are-ta.
John-NOM Emily-AGENT big-room-LOC alone-left
satisfied-until cry-CAUS-PASS-PAST
('Being left in a big room alone, John was allowed to cry by Emily
until he was exhausted.')

However, in the case of unaccusative verbs, the situation seems to be more complicated. Kageyama used only the short causative in his examples (p. 61). That is, he ignored long causative cases, which should be regarded as real causativization in a strict sense rather than transitivization. Let us consider an example of causative-passivization for an unaccusative verb in (35). As Kageyama described, the sentence becomes unacceptable (35a). However, again with the help of a contextual adverbial phrase, it appears to be an acceptable sentence[28] . The sentence (35c) is an example of a direct passivization from (35b). Here, we notice an analogous situation between causativization (30c) and passivization (35c), for the unaccusative verb.

[28] Actually, Kageyama also acknowledged in his note that there are speakers who accept these unaccusative causative-passive sentences. He suggested a possible NP-movement of the underlying object (p.73), though he did not give the linguistic nature of the movement.

(35) a. ?*Yasai-ga KUSAR-{as, ase}-[r]are-ta. (modified from Kageyama, 1993, p.61)
vegetable-NOM rot-CAUS-PASS-PAST
('The vegetable was made to rot.')

b. Kare-ha Sono-Zikken-de 10Kg-no Yasai-o KUSAR-{as, ase}-ta.
he-TOP DET-experiment-LOC 10-kilogram-GEN vegetable-ACC rot-CAUS-PAST
('He made ten kilograms of vegetables rot in the experiment.')

c. Sono-Zikken-de 10Kg-no Yasai-ga KUSAR-{as, ?ase}-[r]are-ta.
DET-experiment-LOC 10-kilogram-GEN vegetable-NOM rot-CAUS-PASS-PAST
('Ten kilograms of vegetables were made to be rotten in the experiment.')

Furthermore, there is another, probably more serious, counter-example to Kageyama's assertion. In our analysis of causativization, we showed a characteristic behavior of type (F) unaccusative with Vi root. If we apply passivization to this type of verb, we get a perfectly natural sentence, as in (36).

(36) a. Omocha-no Booto-ga Otokonoko-niyotte SIZUM-e-[r]are-ta.
toy-GEN boat-NOM boy-by sink-TRANS-PASS-PAST
('The toy boat was sunk by a boy.')

b. Omocha-no Booto-ga Otokonoko-niyotte SIZUM-ϕ-{as, ase}-[r]are-ta.
toy-GEN boat-NOM boy-by sink-INTRANS-CAUS-PASS-PAST
('The toy boat was forced to sink by a boy.')

An important point we observed is that passivization of an unaccusative short causative sentence is possible if the verb type is appropriate, or the sentence is given in the proper context. We do not need, for our purpose, to look for a linguistic justification of this complicated situation. In a conclusion to our discussion about passivization, from our practical point of view, on one hand the applicability of passivization is very impressive, and on the other the production seems to be controllable by carefully associating the verb types.

F. Other Formations

In previous sections, we have introduced two major realizations of syntactical valency control strategies, i.e., causativization and passivization. There are many other possible syntactical realizations of grammatical paraphrasing. In the following subsections, we will separately describe two supplemental strategies, i.e., potentialization and verbal nouns.

1. Potentialization

In Japanese, there are several forms of potential constructions as shown in (37a-d).

(37) a. Hinan-ga Jisin-notoki Tor-e-ru Saizen-no Miti-da.
evacuation-NOM earthquake.when choose-POTN-ADJ best way
('Evacuation is the best way one can choose when an earthquake occurs.')

b. Hinan-ga Jisin-notoki Kangae-rare(or -re)-ru Saizen-no Miti-da.
evacuation-NOM earthquake.when consider-POTN-ADJ best way
('Evacuation is the best way one can consider when an earthquake occurs.')

c. Hinan-ga Jisin-notoki Sentaku-deki-ru Saizen-no Miti-da.
evacuation-NOM earthquake.when choose-POTN-ADJ best way
('Evacuation is the best way one can choose when an earthquake occurs.')

d. Hinan-ga Jisin-notoki Tor-kotoga.deki-ru Saizen-no Miti-da.
evacuation-NOM earthquake.when choose-POTN-ADJ best way
('Evacuation is the best way one can choose when an earthquake occurs.')

e. Hinan-ga Jisin-notoki Tor-i-e(or -u)-ru Saizen-no Miti-da.
evacuation-NOM earthquake.when choose-ADV-POTN-ADJ best way
('Evacuation is the best way one can choose when an earthquake occurs.')

When the verb ends in a consonant (37a), the potential formative "-e" is added to the verb, and when it ends with a vowel (37b), the formative "-rare" (in a conversational style, "-re") is used. We already described in this chapter (section II.D) that in some cases, this "-e" causes valency ambiguity with lexical entities. Also, "-rare" is ambiguous with the passive formative, as well. These ambiguities can cause problems with the accuracy of indexing. We will say more about these problems later (subsection III.B.6c).

A formative, -deki is a potential form of the copula "-s," and is found attached to a verbal noun as found in (37c), but when combined with a formal noun "-koto," a more complicated form -kotoga.deki can attach to a verb as shown in (37d). The last form, "-e" (or "-u") can attach to an already conjugated (adverbial) form as seen in (37e), and is used on rather rare occasions, and gives a very formal impression and classical literary flavor.

All of the above potential patterns are syntactically realized, and are almost universally applicable unless the verb is stative (e.g., AR 'be (inanimate)' and IR 'be (animate)'). (Potentialization is applicable even after the causativization. However, a potential of a passive sentence is often semantically odd.) In any case, potential constructions do not affect the valency of the verb in a sentence. Thus, potentialization is not a valency control strategy though it is a grammatical paraphrase technique. We add potentialized queries to our experiments as supplements.

2. Verbal Nouns

We already described the verbal noun's two-sided characteristic as verb and noun in Section D in this chapter. Thus, a verbal noun has an argument structure at the level of underlying structure, but is morphologically invariable. Furthermore, the argument of a verbal noun requires its own highly selective semantic class in order to yield very specific lexical semantics to the verbal noun. The effect of VN's lexical semantics on retrieval performance was discussed in Fujii and Croft, 1994.

Although a verbal noun cannot change its own form, and consequently does not have the means of carrying out transitivity alternations, a verbal noun in the predicative construct can make itself causative (38a), passive (38b), or potential (38c) by changing the attached copula's form correspondingly (i.e., to *-s-ase*, *-s-are*, or *-deki*, respectively). In addition to these predicative constructs, a verbal noun can demonstrate the genitive construction in order to make a noun phrase (38d).

(38) a. Boss-ga Keikan-ni Yougisya-o Taiho-s-ase-ta.
boss-NOM police-DAT suspect-ACC arrest-COP-CAUS-PAST
('His boss made the police arrest the suspect.')

b. Hannin-ga Keikan-niyotte Taiho-s-are-ta.
suspect-NOM police-INSTR arrest-COP-PASS-PAST
('The suspect was arrested by the police.')

c. Keikan-ga Hannin-o Taiho-deki-ta.
police-NOM residents-ACC arrest-COP(POTN)-PAST
('The police could arrest the suspect.')

d. Hannin-no Taiho
suspect-GEN arrest
('arrest of the suspect')

From a "weak" lexicalism framework, the structure of a verbal noun plus copula "s" was analyzed as an incorporated structure by Kageyama (1993), and the corresponding relationship between the verbal noun genitive phrase and the verbal compound was analyzed by Kageyama and Shibatani (1989). Thus, incorporation or compounding are the primary grammatical devices for verbal nouns to realize variable grammatical relationships, so they are quite different from derivational and inflectional suffixes for verbs. Nevertheless, verbal nouns display a (syntactical) valency control mechanism with different grammatical devices (although they lack a lexical means for valency alternations), and we add verbal noun queries to our experiments as supplements.

G. Summary of Linguistic Knowledge Model of Information Retrieval

In this chapter, we introduced the grammatical paraphrase model of query formulation as a general linguistic retrieval technique. We developed a framework of valency control strategy as its specific realization. To implement various valency control strategies in a real language, we examined the Japanese verbal system, especially transitivity alternation in morphology, and causative and passive constructions in syntax. There, we found two important dichotomies, i.e., valency dichotomy and linguistic module dichotomy, which can be regarded as a very basic conceptual framework of linguistic information retrieval. The former is a contrast of bivalent query construction versus monovalent query construction. The later contrasts lexical method versus syntactic method. In contrast to English, Japanese provides a well suited place to test these two dichotomies. While the dichotomous problem setting implies two mutually exclusive concepts, it revealed that this attitude is too simplistic and naive when we analyzed the details of Japanese morphology. Although we have to maintain conceptual modularity in the process of model building as a scientific venture, we witnessed many intriguing situations, such as intransitivity as a concept composed from unaccusativity and unergativity, which originate from D-structure; Japanese morphological doublet patterns, which structurally refer to the transitive and intransitive categories of the root; and morphological relatedness between lexical "short" suffixes and syntactical "long" formatives.

In the next chapters, we will start constructing hypotheses based on the above conceptual dichotomous framework. We will then address other methodological issues such as text indexing methods, experimental corpus, and query formulation methods.

CHAPTER III

METHODOLOGY

In this chapter, we will first define our hypotheses concerning retrieval performance in terms of valency control query strategies based on the theoretical background discussed in the previous chapter, so that we can then conduct experiments to prove (or disprove) these hypotheses. In the first section, we will describe various hypotheses deriving from our basic experimental design. In the following sections, we will describe the various technical methodologies to be implemented in the experiments. First, we will present indexing techniques, such as the definition and creation of various linguistic indexing cues by using certain natural language techniques. Second, our experimental conditions, such as the characteristics of our test corpus and the description of information need (i.e., a narrative (natural language) representation of the query), will be described. Finally, the methodology of query formulation will be presented. This final section will include two optimization techniques for use in the selection of verb-noun combinations, i.e., the relevance feedback method and the automatic method.

A. Basic Hypotheses and Experimental Design

Deriving or predicting a significant characteristic of the target system from the theoretical model, then defining the character of that target mechanism as a general hypothesis, is a crucial initial step in any scientific experimentation. In our case, the linguistic issues of the valency control mechanism in Japanese discussed in the previous chapter form such an underlying model for our experimentation. We should test the validity of the model through hypotheses, and ultimately build an effective retrieval technology.

In the course of our theoretical discussion in Chapter II, the issues were constantly linked to two general problems - the "valency" problem of verbal constructions and the "interface" problem of linguistic modules. These two problem settings lead to two interlinked dichotomies - the valency dichotomy and the linguistic module dichotomy. The focus of this research is to find effective retrieval domains within this contingency.

The valency dichotomy, i.e., bivalency versus monovalency, is realized by transitivity versus intransitivity in lexicon, or by causativization versus passivization in syntax. We will merge the lexical and syntactical correspondents and examine them as a combined measure.

We can analyze the linguistic module dichotomy in a similar fashion. In our case, this dichotomy consists of lexicon and syntax. This contrast is realized by the transitive versus causative under bivalency, and also by the intransitive versus passive under monovalency. We can make a corresponding combined measure for these two areas as well. Adding potentialization and verbal nouns as supplemental categories, we designed the following arrangement of factors for our experimentation. Thus, Table 6 shows the scheme of our basic experimental design.

Table 6. Basic experimental design.

<baseline>	Syntax	Lexical	Combined	VN
Monovalent	(1)	(2)	(3)	(10)
Bivalent	(4)	(5)	(6)	(11)
Combined	(7)	(8)	---	(12)
Potential	(9)	---	---	(13)
Genitive	---	---	---	(14)

The valency dichotomy is translated into the following hypotheses: (Note that in following hypotheses, we simply picked up one side of the proposition as an assumable better choice. But, in the practice of the experimentation, we might find the opposite direction more suitable.)

- Hypothesis-TI [Transitive (5) vs. Intransitive (2)]: The transitive strategy is more effective than the intransitive strategy.
- Hypothesis-CP [Causative (4) vs. Passive (1)]: The causative strategy is more effective than the passive strategy.
- Hypothesis-BM [(3) vs. (6)]: The bivalent (i.e., combination of transitive and causative) strategy is more effective than the monovalent (i.e., combination of intransitive and passive) strategy.

Potentialization (9) is also a syntactical construction, however it does not change the valency neither increased nor decreased. Therefore, we did not create specific individual valency dichotomy hypotheses for this category, but we will observe its performance in our experiments. On the other hand, verbal nouns can be both causativized and passivized. Therefore, we construct the following syntactical valency hypothesis for verbal nouns:

- Hypothesis-VNCP: [VN Causative (11) vs. VN Passive (10)]: Using verbal nouns, the causative strategy is more effective than the passive strategy.

For the linguistic module dichotomy, we propose the following hypotheses:

- Hypothesis-TC [Transitive (5) vs. Causative (4)]: The transitive strategy is more effective than the causative strategy.
- Hypothesis-IP [Intransitive (2) vs. Passive (1)]: The intransitive

strategy is more effective than the passive strategy.

- Hypothesis-LS [(7) vs. (8)]: The lexical (i.e., combination of transitive and intransitive) strategy is more effective than the syntactic (i.e., combination of causative and passive) strategy.

Based on the above, we may want to compare the behavioral profiles of the valency dichotomy and the linguistic module dichotomy:

- Hypothesis-VL [Valency vs. Linguistic module]: The performance of the valency dichotomy has a different profile (variance, etc.) from that of the linguistic module dichotomy.

Verbal nouns are grammatically and semantically different significantly from verbs. So, expectedly, do their strategic effects, and we consider the following hypotheses:

- Hypothesis-VVN [Verb (7, 8) vs. VN (12)]: The verbal strategy is more effective than the verbal noun strategy.

Finally, if the majority of individual techniques give considerably positive results, we may affirm the following general assertions:

- Hypothesis-G: The valency control technique is effective for retrieval.
- Hypothesis-MONO: The monovalent strategy is significantly effective.
- Hypothesis-BI: The bivalent strategy is significantly effective.
- Hypothesis-LEX: The lexical strategy is significantly effective.
- Hypothesis-SYN: The syntactic strategy is significantly effective.
- Hypothesis-VN: The verbal noun strategy is significantly effective.

Before we close this discussion, note the following two points: (1) Each individual judgment depends on implementation, such as the selection of operators, the method of combining two distinct strategies, etc. We may be able to prove the existence of a specific effective strategy if we can find one. But, if we can not, it does not mean the denial of the future development of an effective technique. (2) The judgment of each hypothesis depends on how the data was measured. Typically this is addressed in relation to the model of retrieval effectiveness measurements, such as recall and precision. For example, given a precision-recall curve, we may calculate the average precision over the total range of recall in order to show the overall performance of both precision and recall, or we may examine precision only in the low recall region in order to see the solo effect of precision. In any case, these are inherent general problems of experimental methodology in IR research. To support and realize these hypotheses, we will develop our experimental methodology in the following sections.

B. Text Indexing

Text indexing is a process to encode texts in a document collection into retrievable forms that are stored in a database. The amount of indexed documents may be very large, as many as thousands to millions (e.g., our test collection has 152,650 documents), so indexing involves some very heavy computing. We should index all sufficiently necessary *index terms* and *retrieval cues* in order to be able to accept any anticipated queries.

Consequently, we have two types of indexing problems. The first is to know which text elements (known as *keywords*) should be extracted from texts as index terms. The second is how to identify which kind of retrieval cues we can add and associate to the text elements to improve the retrieval effectiveness. These two indexing tasks are realized in the following basic indexing steps: (i) Identify each sentence to be analyzed; (ii) Identify words by segmenting the sentence; (iii) Analyze the words morphologically; (iv) Analyze the sentence syntactically; (v) Select the keywords in the text as *primary indexing terms*; (vi) Determine the retrieval cues from the linguistic analysis (of (3) and (4)) as *secondary indexing terms*. In an example of the indexing process (39), (39a) is the original text, (39b) is an intermediate result from step (i) to (iii) to identify the characteristics of every word, and (39c) is the final result of indexing after the selective evaluation of words (by (iv), (v), and (vi)).

(39) An example of indexing process:

```
a. <Original Text>
朝鮮半島の統一が日本経済に及ぼす影響では
日韓経営者の見方は対照的である。
(Concerning the effects of the unification of the Korean peninsula with the Japanese ec
the views of Japanese and Korean business executives are contrary to each other.)
b. <Segmentation Result> (normalized)
朝鮮 N  *
半島 POSTFIX  *
の POSTPOS  "ÇÃ @GEN/#NOM/#ACC"
統一  VN  統一  * VN=Vti=Vti=2/4={@NOM/@ACC/@0}=*=@GOAL(*)=@WITH(*)"
が POSTPOS  が  #NOM"
日本 N  *
経済 N  *
に POSTPOS  に  #GOAL/@CAUSE"
及ぼす V  及ぶ  @CAUS-s V=Vd=Vis=3/3={@NOM/@ACC/@DAT}=*"
影響  VN  影響  * VN=Vt=Vt0=2/2={@NOM/@0/@DAT}=*"
で POSTPOS  で  #CAUSE/#AT"
は POSTPOS  は  #TOPIC/#NOM/#ACC"
日韓 N  *
経営  VN  経営  * VN=Vt=Vt0=2/2={@NOM/@ACC/@0}=*"
者 N  *
の POSTPOS  の  @GEN/#NOM/#ACC"
見方 N  *
は POSTPOS  は  #TOPIC/#NOM/#ACC"
対照  VN  対照  * VN=Vt=Vt0=2/3={@NOM/@ACC/@0}=*=@GOAL(+    と)"
的 AN  *
```

```
だ  ASSERT  *
。  SPECIAL  *

c. <Generated Indexing Terms>  (separated by "|")
朝鮮 |半島  | @GEN_ACC |統一  | @SUBJ |日本 |経済  | #GOAL/@CAUSE |
及ぼす_@ORG | 及ぶ_@NF | @CAUS-s |影響  | #CAUSE/#AT |日韓  |経営  |
者  | @GEN_ACC |見方  | @SUBJ | @DO|対照 |的 |だ |。  |
@FIL | @FIL | @FIL | @FIL | @FIL | @FIL | @FIL | @FIL | @FIL | @FIL | @FIL |
@FIL | @FIL | @FIL | @FIL | @FIL | @FIL | @FIL | @FIL | @FIL | @FIL | @FIL |
@FIL | @FIL | @FIL | @FIL | @FIL | @FIL
```

Although steps (i) to (iv) describe the basic flow of the processes, they do not necessarily proceed in strictly sequence, and adjacent processes may occur back and forth depending on the technique. In the following subsections, we will discuss the aspects of the indexing process in the order of these six stages, and occasionally refer to an example of each corresponding technique in (33).

1. Sentence Identification

Sentences have two significant and practical reasons to be identified as a unit of syntactic analysis. One concerns the specification of the sentence's grammatical pattern as a query strategy. In our case, a combination of a transitivity verb type and a syntactic construction pattern (i.e., passive, causative, or potential) is translated into an indexing term pattern as a query. This topic will be discussed in the Section C, which concerns the query construction method used in our experiments.

The other reason concerns the taming of what we call the *sentence effect*. Thus, when we specify some grammatical pattern of keywords in a query, the effect of such a pattern on retrieval performance includes not only the effect of its special keyword arrangement, but also the general effect of a *sentence* itself as a single general unit rather than as a simple collection of the keywords. Thus, in order to measure the pure effect of specific grammatical pattern, we should not only discount the contribution of every keyword, but also remove the performance effect of a simple word pattern of a plain active sentence (i.e., sentence effect).[29] For example, if we want to determine the effect of passivization, we have to set a query with an active sentence pattern as the baseline, which would yield different results from the performance of a query with the keywords in random order. (See also III.D.2 about the "three-level query structures.")

Technically speaking, the difficulty of sentence identification by means of the sentence boundaries must differ in different languages and writing systems. In English, a period is a strong boundary clue, but there are special cases where the same mark is used as a decimal point in a number, or an acronym indication. Although old style Japanese text (before 1950), as Inoue (1984) pointed out, lacked systematic punctuation rules, modern Japanese texts have a special symbol "ÅB" (the Kuten) which indicates a sentence end. The role of the Kuten as a sentence terminator is exclusive, so the separation of a sentence in Japanese text is a much easier task than in English or other western languages.

Because most IR systems, including INQUERY, which was used in our experiments, do not provide the sentence as a predefined entity (i.e., "primitive") in the system, their operators are not workable as sentence-wise functions. So, for example, if we specify a grammatical sequence of keywords (e.g., a subject-verb order) by a proximity operator which specifies the sequence with its "window" size, we have to secure the sentence boundaries

[29] There is an analogy between this idea of sentence effect exclusion in our retrieval experimental design and the assumption of underlying structures in the theory of generative grammar. In this sense, keywords in a query corresponds to a set of lexical entries for the process of lexical insertion.

so as not to run beyond the sentence. To implement this operational constraint, our indexing program adds a series of "fillings" between two adjacent sentences as seen in (33c) - twenty eight filling items ("@FIL") are added after every Kuten symbol. This filling length was determined as being a size equal to the average sentence length in our Japanese test collection. Out of 10,000 randomly sampled sentences in our collection, the average length in words was 27.9 , and the standard deviation was 14.1 . Thus, if we specify a proximity operator with a window size (or less)[30] which is less than the length of this filling sequence, the operator will never cross the preceding or succeeding sentence boundary. More about the usage of retrieval operators in our experimental queries will be discussed later in this chapter (see Section C).

2. Word Identification

Before moving to the issue of word identification, let us first consider what kind of information is an appropriate and effective unit for indexing. We define two distinct indexing levels, the morpheme level and the word level. In Japanese, a kanji character is regarded as a morpheme, a minimum meaningful element in the composition of a word. Using this characteristic of kanji, we compared the performance of morpheme-level retrieval and word-level retrieval in our previous experiments (Fujii & Croft, 1991). According to the results, if we properly formulate a query (i.e., *post-coordination*) which is made of separated kanji characters, using the #phrase operator of INQUERY, for example, then *morpheme-based retrieval* can perform better than *word-based retrieval*. Morpheme-based retrieval showed

[30] It is reasonable to assume that the distance between the verb and its argument (especially the object) in a sentence is much shorter than the sentence length. Although the number of filling items and operator window size are empirically determined, our experiments worked well as seen in the next chapter.

an average improvement of 9% on average. The improvement was most noticeable at the middle to high recall level. Thus, this technique is regarded as a recall enhancement device. Furthermore, we demonstrated that if we indexed both kanji and words together, performance was improved at all recall levels (14% improvement on average) (Fujii & Croft, 1993).

One problem of character-level indexing is the difficulty in applying linguistic techniques to the indexed terms. In other words, once a word is decomposed into characters, the identity of the word as a syntactical or morphological unit can be lost, or at least we need an extra effort to recover the unity. Furthermore, in contrast to the enhancement of recall performance by character-level indexing, our grammatical query strategies will probably result in precision enhancement because they filter out less relevant text by matching the grammatical patterns. Therefore, character-based and word-based indexing are not opposed methods, but are complementary methods. Thus, in this study, we simply explore the potential advantages of linguistic methods based on word-based indexing.

Now we return to the original question. How can we identify and separate every word in a sentence? When a writing system provides a way to identify words by punctuation as in English, which places a space between two words as a delimiter, this task is easy. However, in a language like Japanese, in which text is written without space between words, this process, "(sentence) *segmentation,*" is not a trivial task. There are various segmentation algorithms, such as the longest-matching method, etc., as Tanaka (1989, pp. 133-153) described.

A segmentation program called Juman (Matsumoto, et al., 1991), which was used in our experiments principally adopts both the longest-matching heuristic and the connection table method, which utilize all valid sequences of grammatical categories (i.e., parts-of-speech or sub-categories) to be assigned. Juman scans the sentence from right to left and looks up the longest string that matches a dictionary entry, but which is also a legitimate sequence of grammatical categories with proper inflectional endings. Thus, Juman not only works as a "sentence segmenter," but also functions as a "part-of-speech tagger" to identify grammatical categories and related lexical information. In Juman, the user can build his/her own dictionaries and grammatical relationships. The contents and the utilization of these types of information will be described in the following discussions.

3. Morphological Analysis

Morphological analysis of word structure is generally divided into two types, inflectional analysis and derivational analysis. Inflectional analysis handles inflectional suffixation. Derivational analysis consists of two major components, which target derivational affixation, for our purpose especially suffixation, and compounding. As a morphological analyzer, Juman was designed to combine access to a permanent lexicon (i.e., a dictionary), and identification of inflectional affix patterns. However, it lacks a specialized module for derivation. Thus, it recognizes two characteristic items: one is the lexical entries of a dictionary, which are permanent and static, and the other is the elements of a verb or adjective's inflectional paradigm, which carry out the syntactical functions and which are generally assumed to be regular and productive.

Regarding inflectional changes, Juman identifies the root and the inflectional ending together. The indexing program which works together with Juman converts the inflected form into the infinitive ("Shuushi-kei") form to normalize the paradigmatic variations. In (33b, line 9), Juman normalized the adjectival form of a verb OYOB-os ("及ぼす", 'influence') into the infinitive form, which happens to be identical to the original adjectival form in this example. (Of course, this is not the case in most cases.) This normalization process is similar to the conventional stemming procedure, which removes the attached endings to index only the common "stem," but there are several differences between these two operations. In general, the former tends to be applicable to generic domains, but the method is often ad hoc; the latter may be more theoretically identifiable, but at the same time its applicability might be more limited to specific categories. Specifically, the normalization operation is not necessarily limited to the stemming of suffixes, but potentially allows one to change the form quite greatly based on the normalization principle. In this sense, normalization is a realization of theoretical transformation, as in our transitivity alternation model. Furthermore, the output of the stemming process usually becomes immediate material for indexing, but a normalized string may be used as material for other stages of indexing process. Indeed, in our experiments, the results of inflectional normalization were further converted into a more basic form in transitivity alternation, as we will describe later in this section.

Next, let us consider the case of derivational suffixation. Although we discussed and viewed transitivity alternation as a productive and systematic derivational process in Chapter II, and it therefore performs one of the key roles in our linguistic retrieval experiments, the mechanism is not implemented in Juman in a functionally productive way. In other words, unchangeable entries in a dictionary are treated as if they were the end results of already completed derivational production. Consequently, in Juman's verb dictionary, the coverage of manually collected verb doublets is often incomplete, and worse, these doublets lack information on transitive-intransitive associations such as common roots and their derivational types. To correct this limitation of Juman, we implemented a computer program to automatically create a verb dictionary that includes all transitive alternative forms derived from the entries of the "root dictionary." This program produced 13,849 verbs from the 3,271 verb roots in our experimental root dictionary. In (33b, line 9), the basic intransitive form OYOB-ϕ ("及ぶ", 'reach') of the transitive form OYOB-os ("及ぼす", 'influence'), which had been normalized into the infinitive form before, is recognized, and the verbal type, short causative (@CAUS-s), is annotated. Based on these pieces of information, both the original transitive form (及ぼす_@ORG), the final basic intransitive form (及ぶ_@NF), and the marker (@CAUSE-s) are written as indexing terms in (33c).

Next, let us discuss the case of compounds in derivation. Although compounding is not a main focus in our experiments, it requires some consideration. Unlike compounds in English, which are either "open" (that is, written with spaces between the component words, e.g., 'text critique'), hyphenated (e.g., 'morpho-syntactic'), or "closed" (that is, written with neither spaces nor hyphens between the components, e.g., 'textbook'), Japanese compounds must be identified completely without the assistance of punctuation clues. For example, Juman identifies a segmented string as a compound based simply on the dictionary entry to be matched, so that there is no further consideration of whether a compound is divided into elements or not. Thus, if the dictionary contains compounds in addition to their elements, then Juman will extract only the compound as a single unit because of its longest-string-matching heuristics.

Here, we find different segmentation guidelines between dictionaries constructed for the purpose of retrieval indexing, and those for other NLP applications such as text understanding or machine translation: for retrieval indexing, segmentation must be more element-driven. For example, to make 'textbook' retrievable by querying 'book' (to improve recall), 'text' and 'book' should be indexed separately and be made retrievable as "text+book" by a post-coordinating specification (to improve precision), so that 'book' can be retrieved (supposedly with a smaller term weight to lower the rank) as well as 'textbook.' In other NLP applications 'textbook' must be a single unique concept, and the understanding of the semantics may do the rest of the business. Consequently, the principle of compound handling for retrieval indexing is that of breaking-up (i.e., word segmentation) into constituent elements, unless the compound is totally idiosyncratic,[31] such as 'White House.'

Since our concern is verb relationships, the next natural question is whether the above guiding principle of compound indexing should be made applicable to a "verbal compound," or not. Verbal compounds are much more highly grammatically constrained than arbitrary associative "primary compounds" (Spencer, 1991), and they typically have a N+V (e.g., 'truck

[31] If lexical ambiguity of a compound requires the condition, or in order to capture a wider range of retrieval applicability, both the compound and the elements may be indexed together.

driver'[32] or Tenki-Yohou 'weather forecast') or V+V (e.g., Tobi-Kom (=jump-enter = 'jump-in')) pattern. The former can be seen in both English and Japanese, whereas the latter is very rare in English but is often found in (original non-Sino) Japanese verbs. When the pattern is N+V, we simply separate N and V. In any case, separation of a verbal element from a compound should be beneficial for linguistic retrieval, which utilizes the verbal function. And considering the significance of valency control as our purpose in retrieval, let us concentrate on V+V type compounds in the following discussion.

Kageyama (1993) analyzed Japanese verbal compounds intensively, arguing that there are two distinct formations of such compounds, that is, lexical compounds and syntactic compounds. Syntactic compounds are generated in the syntactic component of the linguistic architecture, so the meaning is transparent, and dissociating the elements should not cause any problem for retrieval. Moreover, since the majority of heads in syntactic compounds are merely aspectual, like *Oeru* 'finish' as seen in *Tabe-Oeru* 'finish eating,' extracting the modifier element of a syntactic compound, in this example *Tabe* 'eat,' is a very important task for retrieval.

On the other hand, the situation of lexical compounds is more complicated. Kageyama (1983) showed that some lexical verbal compounds demonstrate his "*Transitivity Concordance Principle*" (TCP), which means that an unaccusative element can be combined only with another unaccusative element, and cannot be coupled with a transitive or unergative

[32] This is a case of $[N + V + er]_N$. However, there are very few $[N\ V]_V$ compounds in English, and this arises a controversial point whether *track driver* should be analyzed as $[[\text{truck drive}]_V\ \text{-er}]_N$ or $[[\text{truck}]_N\ [\text{drive -er}]_N\ ]$. Here, we simply assume that *truck* satisfies the internal argument at the First-Sister position (Roeper and Siegel, 1972) of *drive* . These arguments are summarized in Carstairs-McCarthy (1992, pp. 108-119).

element regardless of whether it is a modifier or a head. Similarly, a transitive (or unergative) element can associate with only a transitive or an unergative element. TCP was introduced as a mechanism of the integration of two argument structures of elements. However, as Kageyama observed, in another group of verbal compounds, the TCP seems not to hold, by making a combination of a transitive and an unaccusative such as *Sui-Tuk* ('suck-and-attach(vi) (by mouth etc.)'), *Oti-Tuke* ('fall-and-attach(vt)' = 'settle down'), and *Nagare-Das* ('flow-and-take.out' = 'flow out'). Kageyama gave some explanations of these cases using several linguistic mechanisms, such as the double coding (i.e., polysemous) of an unaccusative and unergative on a single element, back-formation (or reanalysis), and formation within lexical conceptual structures rather than within argument structures. Finally, the simple identification of the head transitivity does not guarantee the compound's transitivity. In (40a, b), the heads of the verbal compounds differ as to their transitivity, but the meanings of two compounds (consequently the sentences) are almost the same.

(40) a. Kodomo-ga Ie-no Soto-ni Tobi-De-ta.
child-NOM house-GEN outside-DAT jump-exit(vi)-PAST
('The child jumped to the outside of the house.')

b. Kodomo-ga Ie-no Soto-ni Tobi-Das-ta. (s-t=>sit)
child-NOM house-GEN outside-DAT jump-exit(vt)-PAST
('The child jumped to the outside of the house.')

It may be possible to make a sophisticated mechanism of verbal compounds based on Kageyama's model, but in my judgment, at least in this series of experimentations, the determination of compound transitivity is not necessary for retrieval indexing if we index every compound element and normalize its transitivity alternation to absorb the fluctuation of transitivity as done for syntactical compounds. With this treatment, all required arguments of verbal compound elements will be captured, with no serious retrieval drawbacks for compound transitivity handling.

4. Syntactic Analysis

There are various kinds and levels of syntactic analysis and techniques. Although this is a very rich field of research in natural language processing, for our purpose I want to take only two issues into consideration. One is to identify the subject and object (if the verb is transitive) of a sentence. The other is to detect passive, causative and potential constructions.

The identification of subjects and objects needs, in general (especially when the target language shows weak morphology, e.g., English), a full syntactic analysis through parsing of the sentence. Furthermore, if the subject is missing from a sentence (especially in a "subject-drop" language like Japanese), deeper analysis is required. To avoid this computationally expensive process, we adopt a shallower, but much easier method. Thus, we skip the process of full parsing, but mark the characteristic postpositions as secondary indexing terms. In Japanese, postposition "*ga*" is used in most cases as a subject (nominative) marker. There is an example of this subject marker (@SUBJ) in (39c). Postposition "*wa*" is used as a topic marker, and is also a strong cue of a possible subject item. However, there is also a chance that it indicates an object item. Therefore, we marked both subject and object indicators at the same position of this topic marker to allow either case to be retrieved. (However, we threw out the topic marker when it was compounded after another postposition such as "-de+wa" (#CAUSE/AT+ #TOPIC) in (39b, line 11 and 12) before (39c).) Furthermore, postposition "*o*" strongly indicates a direct object (as an accusative) of the verb. Examples of these postpositions are shown in (41).

(41) a. Sono Saru-ga/-wa Ori-o KOWA-s-ta. (s-t=>sit)
the macaque-NOM/-TOP cage-ACC break-TRANS-PAST
('The macaque broke the cage.')

b. Sono Saru-wa Ori-wa KOWA-s-nai. (s-n=>san)
the macaque-TOP cage-TOP break-TRANS-NEG
('The macaque doesn't break the cage.)

We did not handle the problem of subject drop in our experiments. However, in formal writing style, such as our newspaper articles, subject drop is likely to be less frequent than in more informal text or speech, in order to make the flow of logic clearer to readers.

Next, concerning the detection of the syntactic constructions of the passive, causative and potential, the situation is similar to the above tasks of subject and object identification. Again, in general we need a full constituent analysis to achieve an impeccable result, but some languages, including Japanese, have postverbal marking to indicate these constructions, which makes the judgment possible in local morphological processing. In Japanese, as we discussed in Section II(E), the suffix (or auxiliary) "-ase" is used for the (long) causative, "-are" for both passive and potential. Thus, we place their secondary indexing markers where these suffixes appear in the text. Since these morphemes as syntactical affixes are morphologically stable in modern Japanese, the technique we adopted here must be robust. Similar morphological techniques should be applicable to various languages if they have specific morphological devices for causativization and passivization.

5. Selection of Keywords as Primary Indexing

Not all words identified by a segmenter (e.g., Juman) need to be indexed, and only selected terms which are expected to contribute to the retrieval are. This elimination process is commonly referred as stopword removal. Although the utility of stopwords has been intensively discussed in IR literatures, it mainly affects space efficiency, and has only a slight impact on retrieval effectiveness. So, in this paper, instead of looking at the concept of stopword, we instead inquire into the selecting of primary indexing terms.

To achieve the goal of our experiments, it is makes sense to divide the words as index terms into three groups rather than the standard two groups keywords and stopwords. Thus, namely, we count: i) *stopwords*, ii) *peripheral keywords*, and iii) *frame keywords*. The concept of stopword in this new term framework is derivationally the same as the standard notion of stopwords, but operationally different. In the traditional view, stopwords consist of grammatically varied words such as pronouns, adpositions (i.e., prepositions and postpositions), adverbs, and even some nouns and verbs, because they are chosen mainly by the frequency of the word. However, under circumstances like our grammatical IR experiments, stopword elimination must proceed very carefully in order not to taint the basic grammatical framework and conceptual relationships in a sentence. Therefore, we either have to carry out no stopword elimination at all, or carry out the elimination after some morphological/syntactical analysis. In our experiments, we did not eliminate any words.

Keywords are divided into peripheral and frame keywords in our experimental indexing system. The frame keyword is defined as a constituent of an argument structure on which the valency control operation is executed;

peripheral keywords are the rest of the words outside of stopwords and frame keywords. Frame keywords principally include the main verb and its subject and objects; typical peripheral keywords are nominals found in an adjunct phrase that does not participate in the argument structure. Although peripheral terms will not change the performance of the valence control, they should be expressed in the query as a part of the baseline. On the other hand, query modification by a valence strategy that contains frame terms should improve the retrieval apart from the baseline. A significant difference of our grammatical stopword/keyword system from the traditional system is that the membership of a word is not fixed as in a static list of words, but may change sentence by sentence depending on the structure. The determination of peripheral keywords will be addressed in the last section of this chapter, which concerns creating the baseline; the arrangement of frame keywords for retrieval strategy will be discussed in the next chapter, which includes more specification of query organization.

6. Marking Linguistic Cues as Secondary Indexing

In contrast to primary indexing, secondary indexing here means that the indexing terms are not extracted as keywords directly from the text, but are derived from the results of some further analysis. After such linguistic analysis, extracted grammatical features (i.e., cues or markers) are properly embedded in already indexed text keywords, and stored in a database linked with their location information (in the text) as secondary indexing terms. Thus, linguistic secondary index terms are the main vehicle with which to carry out our valency control strategy to improve the retrieval performance.

Before describing the implementation of grammatical clues as secondary indexing terms in our experiments, let us consider the issue of the trade-off and balance relationship between pre-coordination and post-coordination (Van Rijsbergen, 1979, pp. 22-23) in retrieval. Thus, if we analyze the text more deeply and put more knowledge into the database, at the time of text indexing, i.e., at the pre-coordination stage prior to all query processes taking place, query processing will become functionally simpler, and the load put on query formulation will be lighter. On the other hand, less pre-coordination processing requires more sophisticated and more costly query operation, i.e., heavier post-coordination. If the retrieval system does not have a sufficient repertoire of retrieval operators to capture and express desired grammatical properties, such text features must be pre-coordinated. For example, if we do not detect and mark the pattern of passive sentences at text indexing time, we have to have an operator to find the passivization pattern at the run-time of the query. So, the balancing of these two coordinations is important in a practical system.

There are three important points in pre-coordination for grammatical relationships. First, there is the problem of the indexing coverage. Thus, it is trivial that we cannot retrieve some aspect of the text if the pre-coordination process lost the information. Consider the sentence (42a):

(42) a. Mary was bitten by the dog.
b. Mary bit the dog.

Assume that we do not mark the passivization, but only the main verb, so that the indexing result will be {'Mary', bite', 'dog'}. In this example, we cannot distinguish (42a) from (42b).

Second, there is the problem of the structural fidelity of indexing. Thus, an indexing method of grammatical aspects may not provide sufficient structural information, and may cause some interference with other elements (potentially by a side effect). Thus, we may not be able to recover the original sentence pattern from the indexed data. Let us take a look at the following sentence (43a):

(43) a. John knows that Mary loves the flower.
b. {John, @SUBJ, Mary, @SUBJ, love flower, @OBJ}

Suppose that the sentence (43a) is indexed as in (43b). Also, assume that we have only the proximity operator (PROX) which is order sensitive in operands, and has a window of size 10. Thus, this system lacks the flexible ability of grammatical pattern matching[33]. Then, if we run a query "PROX(John @SUBJ love flower @OBJ)" to intend to match with a sentence "John loves the flower," it will mistakenly match with (43a).

Third, there is the problem of the complemental relationship between the representation carried out by the pre-coordination and the functionality carried out by the post-coordination. Thus, a certain form of pre-coordination requires a certain functionality of the post-coordination operator to achieve the objective of the indexing method. For example, the indexing terms in the index term representations (43b) for the sentence (43a) require an order-sensitive proximity operator, and if not, we cannot give the query such as above "PROX(John @SUBJ love flower @OBJ)" to the system.

Now let us move to the contents of secondary indexing terms in our experiments. As we have discussed in subsections 2, 3 and 4, the essential tasks of word identification, morphological analysis, and syntactical analysis

[33] Unfortunately, most current text retrieval systems lack such flexible pattern matching capability; so does the INQUERY system used in our experiments.

in Japanese can, for our purpose, be done by JUMAN's segmentation function, and the following four groups of grammatical features must be identified in the indexing process: 1) subject and object of the sentence; 2) transitivity and intransitivity (by identifying the suffixes of short causative, short passive, and short potential); 3) causativization, passivization, and potentialization (by identifying the (long) causative, passive, and potential formatives); 4) genitive relationship to a verbal noun with a noun which is either subjective or objective. Let us discuss these groups one by one.

a. Grammatical Relational (Subject/Object) Markers

The first group of grammatical information is the subject and object of a sentence. By examination of the categories of adjacent elements, the JUMAN algorithm identifies the case-postposition "-ga," separated from the conjunctive use of this formative. After segmentation by JUMAN, the indexing program assigns the nominative case to the "-ga" postposition, and inserts a subject marker (@SUBJ). One significant exception is that the postposition "-ga" may be used for the direct object in a potential sentence such as in (44a):

(44) a. Sono Saru-ni Kago-ga KOWA-s-e-ta.
the macaque-DAT(AGENT) basket-NOM(PAT) break-TRANS-POTN-PAST
('The macaque could break the basket.')

b. Sono Saru-ga Kago-o KOWA-s-e-ta.
the macaque-NOM basket-ACC break-TRANS-POTN-PAST
('The macaque could break the basket.')

As seen in this example, the subject nominal receives "-ni" postposition, which is usually used as dative of the indirect object of a verb, and often substitutes for the standard "-ga" to avoid its double appearance in a sentence.

However, such a pattern is relatively rare in text, and will have little practical impact on retrieval. We randomly sampled 100 instances of postpositional "-ga" from different articles in the text corpus of our experiments, and checked their categories. In this statistical test, 100 cases, i.e., all, were subjective. Therefore, practically speaking it is unlikely to be necessary to specify a thematic item with "-ga" postposition unless it is in a legitimate intransitive or passive sentence. Thus, "Kago-ga KOWA..." (compare sentences in (44)) will most likely lead to either "Kago-ga KOWA-re-ta" 'the basket broke' (intransitive), or "Kago-ga KOWA-s-are-ta" 'the basket was broken' (passive).

Marking the (direct) object of a transitive verb creates a similar situation to the case of the subject. The postposition "-o" has a distinct written form "を" that is almost exclusively used to designate the object of a transitive verb as the accusative case. With some special intransitive verbs such as motion verbs, "-o" indicates the locative phrase such as shown in (45a). Therefore, the passivization of such sentence becomes ungrammatical, as we see in (45b).

(45) a. Onnanoko-ga Ohanabake-o KAKE-tei-ru.
girl-NOM flower-garden-LOC run-PROG-PRES
('A girl is running in the flower garden.')

b. *Ohanabake-ga Onnanoko-niyotte KAKE-rare-tei-ru.
flower-garden-NOM girl-by run-PASS-PROG-PRES
('*The flower garden is run by a girl.')

Again, such a pattern appears infrequently, and is associated only with specific verbs. Even if the ungrammatical passivized pattern of (45b) is specified in a query, such a sentence pattern virtually does not exist in the indexed text. Therefore, no harmful effect will result.

Japanese has another postposition, "-wa," which attaches to a nominal to indicate the topic of the sentence. As Mikami (1960) argued in his popular book that this postposition is a substitute for the nominative "-ga," accusative "-o," or dative "-ni/e" (or locative "-de")[34], they are semantically paraphrasable in (46a-c) in respective order.

(46) a. John-wa HASIR-ta. (r-ta=>tta)
John-TOP run-PAST
('John ran.')
b. Sono Ie-wa Roger-ga TATe-ta.
the house-TOP Roger-NOM build-PAST
('Roger built the house.')
c. Taisyokusha-(e)wa Kuni-ga Nenkin-o SIHARAW-u.
retiree-(GOAL)-TOP government-NOM pension-ACC pay-PRES
('Government pays pensions to retirees.')

Table 7 contains some statistics of major postpositions, of which samples were randomly extracted from 7,547 sentences in our corpus. As Japanese is regarded as a "topic prominent" language, the frequency of "-wa" is significantly high, more than that of nominative "-ga." Proper handling of the topic marker is unquestionably vital for performance improvement by the valency control retrieval.

[34] Although he also claimed "-wa" substitutes for the genitive "-no," it seems to hold only when the relationship is subjective or objective, but not for simple attributive or possessive use. For example, corresponding to (46a-c), the following variations a'-c' are valid:
a'. John-no HASIRi-Kata ('John's way to run')
b'. Sono Ie-no TATe-Kata ('the way to build the house')
c'. Taisyokusha-(e)no SIHARAWi. ('payment to retirees')
However, d' is not possible as an alternative of "Kodomo-no Omotya-kara Namari-ga Kensyutu-s-are-ta" (kid-GEN toy-ABL lead-NOM detect-PASS-PAST = 'lead was detected from the kid's toy '):
d'. *Kodomo-wa Omotya-kara Namari-ga Kensyutu-s-are-ta.

Table 7. Postpositions in Japanese.

Case/Topic Markers	Freq	%
-wa{TOP}	5025	12.1
-ga {NOM}	3539	8.5
-o {ACC}	7492	18.1
-ni (simple) {DAT (to)}	6011	14.5
-ni+other_function (yoru, tuku, toru, etc.) {INSTR, PERCEPT, BENEF, etc.}	667	1.6
-e {DAT (to)}	296	0.7
-kara {ABL (from)}	892	2.2
-yori {ABL (from), than}	32	0.1
-de {INSTR (by), LOC (at)}	3841	9.3
-to {CO-EXPER (with)}	3314	8.0
-no {GEN(of)}	10319	24.9
Total:	41428	100.0
(in # of Sentences	7547)	

Table 8 contains results of a random sampling concerning the use of "-wa" in 100 sample sentences to show possible case substitutions. Examples (47a-c) show the actual examples from the corpus. The large majority (84%) of "-wa" instances are substitutable for by the nominative "-ga," 5% of them are alternatives of the accusative, and only 1% are substitutions for the dative case (indirect object). Thus, only 1% of total "-wa" usage, 12.1%, i.e., 0.12% may participate in argument structures with dative case.

Table 8. Case substitution of topic marker "-wa."

NOM(-ga): 84%, ACC(-o): 5%,
DAT/LOC(-ni/-e/-de): [for indirect object]: 1% / [for adjunct]: 10%

(47) a. Koushou-wa Saishuu-Dankai-o MUKAE-ta
negotiation-TOP final-phase-ACC meet-PAST
('The negotiation reached its final phase.')

b. Shoyuu-Siteiru Toti-wa (Watasi-ga) 20Nen-Mae-ni Shutoku-Sita.
own-ADJ land-TOP (I-NOM) 20-years-ago-DAT acquire-PAST
('I acquired the land 20 years ago.')

c. Housyokuhin-wa Kaku Hyakkaten-tomo Tikara-o IRE-ru Housin-da.
jewelry-TOP each department.store-NOM effort-ACC put-PRES
plan- COP
('Each department store has a plan to put extra effort into jewelry.')

Considering the above points, we decided to index both subject and direct object (but excluding indirect object) markers simultaneously at the place of the "-wa" topic. Therefore, in our indexing scheme, subject and object markers have two possible origins: either directly encoded from a subject/object postposition, or substituted from a topic postposition. (48a-c) is a portion which includes "-wa" (は), and was extracted from the indexing example of (39).

(48)

```
a. 見方は対照的である。

b.
見方  N   *
は  POSTPOS   は  #TOPIC/#NOM/#ACC"
対照  VN  対照   *  VN=Vt=Vt0=2/3={@NOM/@ACC/@0}=*=@GOAL(+     と)"
的  AN   *
だ  ASSERT   *
°   SPECIAL   *

c. 見方  | @SUBJ | @DO 対照 | 的 | だ |
```

b. Lexical (Transitivity/Intransitivity) Markers

All possible lexical variations of transitivity alternation are generated in the verb dictionary as discussed in Section 3. These transitivity alternation patterns in Japanese were presented in Table 3 in Section II-D, so we can count the following six genera of transitive and intransitive verbs (Table 9).

Table 9. Six genera of transitive and intransitive verbs in Japanese.

("Patterns" were encoded in Table 3; WJ shows frequencies in Jacobsen's (1992) list.)

Types	Remarks	Patterns	WJ
TRANSITIVE:			
a. Unmarked Transitive[-ϕ]	Inherent transitive at both root and word level	C, E	42
b. Transitivizing "-e" [-e]	Suffix attached in transitive as well as in intrans.	D, F	133
c. Short Causative [-as]	Suffix attached lexically as well as syntactically	A, B	138
INTRANSITIVE:			
d. Unmarked Intransitive[-ϕ]	Inherent intransitive at both root and word level	A, F	137
e. Intransitivizing "-e" [-e]	Suffix attached in intransitive as well as in trans.	B, E	91
f. Short Passive [-ar]	Not exchangeable with syntactical counterpart	C, D	85

Transitivizing (b.) and intransitivizing (e.) "-e" may be called "short potential" as discussed in Chapter II. If it is represented by the name "potential," it must be able to modify both transitive and intransitive constructions. However, this suffix seldom means "potential," and we cannot uniquely characterize this suffix as a signature of either transitivizing or intransitivizing, but only of lexicalization. Therefore, in our experiments, we integrate a. and b. as transitivization, and d. and e. as intransitivization, in our limited sense. On the other hand, the short causative and the short passive unambiguously define bivalency and monovalency, respectively, though the short passive always occurs in lexicalization, but the short causative is observed not only as lexicalization, but also in syntactical formations.

Thus, one problem of the "short causative" is that it may originate either from the lexicon, or from a syntactic component. In section II-C, we predicted that in formal writing texts, the short causative as a stylistic variant of the long causative occurs infrequently because it is viewed as a colloquial expression in the traditional prescriptive grammar. We collected 100 samples of the short causative patterns from the corpus in our experiments and checked their origin. Only two cases have a clearly syntactic origin, and furthermore, one of these was used in a quote of a person's talk ("Minzoku-tositeno Hokori-o USINAW-as-are-masita" '(I) was forced to give up my ethnic pride'). The other is not exact a case of typical short causative because of its special archaic expression ("Picasso-o-site Tensai-to IW-as(i)-me-ta" '(he) made even Picasso call him genius'), and historical linguistics tells us that the short causative is an earlier syntactic form of the present long causative in Japanese. Therefore, practically speaking, we do not have to carry out further separation of the syntactical short causative from the lexical short causative, and we can treat a short causative as a lexical entity without mistreating these two linguistic phenomena.

Consequently, we regard the short causative (-as), unmarked transitive (-ϕ), and transitivizing short potential (-e) as lexical bivalent forms, and the short passive (-ar), unmarked intransitive (-ϕ), and intransitivizing short potential (-e) as lexical monovalent forms. We indexed four lexical secondary indexing markers, @TRANS ("transitive" for type (a) and (b)), @INTRANS ("intransitive" for type (d) and (e)), @CAUS-s ("short causative" for type (c)), and @PASS-s ("short passive" for type (f)).

c. Syntactic (Causative/Passive/Potential) Markers

In Japanese, the long causative suffix "-(s)ase" and the passive suffix "-(r)are" are syntactic formatives. Although the syntactic construction for causation can be realized with either the long or the short causative suffix, as we discussed in the previous subsection, the syntactic use of the short causative is unusual in formal written text. Hence, the long causative can, practically speaking, be used as the sole indication of causation.

On the other hand, the long passive suffix does not behave so straightforwardly. When the verb ends in a vowel (i.e., a vowel stem verb, or so-called "Ichidan"), the long passive suffix may have either a passive or a potential meaning. And when the verb ends in a consonant (i.e., a consonant stem verb, or so-called "Godan"), the long passive suffix exclusively denotes the passive voice, with no possible interpretation as a potential. In the latter case, the potential is realized by the short potential suffix "-e," which is carried syntactically, not lexically as it is when it is used for making an intransitive verb. Because of this discrepancy of applicable verbs, it is predictable that the majority of long passives are used in passive sentences, and that potentials would be in the minority. In fact, 82% of randomly selected long passive (100) samples from our experimental corpus showed a passive meaning, and only 5% clearly appeared to be potential constructions. The rest (13%) were judged as a kind of "middle" construction, which demonstrates an identical construction to passive sentences in word order or case-postposition, but the semantics includes the potential (e.g., "can be bent," which is a *dispositional* statement rather than *episodic*). For example, the sample sentence (49d) (also see (49g)) has basically the same structure as the passive sentence (49a), but the semantics, i.e., "John can bend the branch" shows a similarity to (49b).

(49) a. Sono Eda-ga John-ni MAGE-rare-ta.
the branch-NOM John-DAT bend-PASS-PAST
('The branch was bent by John.')

b. John-ga Sono Eda-o MAGE-rare-ta.
(John-NOM the branch-ACC bend-POTN-PAST
('John could bend the branch.')

c. John-ga Sono Eda-o MAGE-re-ta.
John-NOM the branch-ACC bend-POTN(informal)-PAST
('John could bend the branch.')

d. Sono Eda-wa John-ni-mo MAGE-rare-ta.
the branch-TOP John-DAT-even bend-POTN-PAST
('The branch was bendable even for John.')

e. Sono Eda-ga John-ni OR-are-ta.
the branch-NOM John-DAT(Agentive) break-PASS-PAST
('The branch was broken by John.')

f. Sono Eda-ga OR-e-ta.
the branch-NOM break-INTR-PAST
('The branch broke.')

g. John-ga Sono Eda-o OR-ϕ-e-ta.
John-NOM the branch-ACC break-TRANS-POTN-PAST
('John was able to break the branch.')

h. Kodomo-ga Domino-o Yuka-ni Umaku Itiretuni NARAB-e-ta.
kid-NOM domino-ACC floor-LOC well straightly line.up-TR-PAST
('The kid lined up the dominoes very well.')

i. Kodomotati-ga Umaku Itiretuni NARAB-ϕ-e-ta. (See also (20).)
kids-NOM well straightly line.up-INTR-POTN-PAST
('The kids could line up very well.')

j. John-ga Sono Eda-o OR-are-ta.
John-NOM the branch-ACC bend-HONOR-PAST
('John broke(Hon.) the branch.')

k. John-ga Mary-ni Sono Eda-o OR-are-ta.
John-NOM Mary-DAT(Agentive) the branch-ACC bend-PASS-PAST
('John had the branch broken by Mary.')

In the author's judgment, the existence of a middle construction does not undermine the overall effect of passive interpretation for effective retrieval. In the statistics we have just described, the passive usage of 82% of the total is a large majority, and more significantly, a middle sentence is, strictly formally speaking, a passive sentence (i.e., the surface subject is still understood as PATIENT), so it is suitable for our grammatical paraphrasing operations in retrieval. Thus, if we strategically specify a passive pattern in a query, we can justifiably claim that it matches with the pattern of middle as well as that of normal passive sentences. Consequently, we can effectively count 95% (82%+13%) of this type of sentence as passive.

In addition to the above discussion, let us briefly examine another two passive-related phenomena. The first one concerns the situation of attaching a long passive suffix "-(r)are" to a consonant verb. In such a case, the suffix works as a honorific expression, as in (49j). Thus, even though both the passive and the honorific exhibit the same verbal morphology, the honorific causes the subject to remain in the same place, in contrast to the passive, in which the original object is placed in the subject position (49e). Therefore, the honorific sentence pattern can potentially jeopardize the retrieval performance of the passivization query strategy. However, again, this negative situation is unlikely because the honorific expression is common in spoken dialogue in a social context, but is not frequently used in an ordinary written text, which tends to be objective and neutral. For example, we found no honorific examples in the frequency statistics just discussed.

The second problem is that there are two different kinds of passivization in Japanese. One is the direct passive, and the other is the indirect passive. The direct passive is the ordinary kind of passive, which has been discussed so far (e.g., (43e)). It is only applicable to transitive verbs. On the other hand, the indirect passive can be applied even to intransitive verbs (unergatives, as we discussed in Section II-e), not only to transitive verbs. As seen in (43k), in the indirect passive structure, a new subject is introduced which is the experiencer of the negative consequence of the action or event expressed in the original sentence. The original subject receives the agentive dative case, and the accusative case of the original object may remain the same. Thus, if the subject marker for the original subject, or the subject marker for the original object in direct passive sentences, is specified in a query, it will not match the indirect passive sentence in the indexed text. However, the indirect passive seems not significant enough to affect retrieval performance. No indirect passive sentence like (49k) appeared in the one hundred samples used for long passive statistics described above.

Concerning the potentialization, we have already described the ambiguity problems when a vowel stem verb is potentialized. That is, a suffix which is identical to the one of long passive (i.e., -rare) is attached. However, as we have already described, the great majority can be treated as passive.

On the other hand, when a verb ends in a consonant, the suffix is identical to (in)transitive "short potential" suffix (i.e., -e) is attached. When the verb type is A (-as/-ϕ), B (-as/-e), C (-ϕ/-ar), or D (-e/-ar), it causes no ambiguity problem. Thus, when the type is A or C, the suffix "-e" is always interpreted as potential. When the type is B or D, "-e" must be interpreted as intransitive or transitive, respectively, because the root cannot take a potential suffix without first being transitivized or intransitivized. In case of the rest, i.e., type E (-ϕ/-e) and type F (-e/-ϕ), in fact, cause ambiguity situations. Thus, in type E, the intransitive form and the potential transitive form have an identical form (e.g., (49f) and (49g)). In type F, the transitive form and the potential intransitive form are superficially identical (e.g., (49h) and (49i)). However, in a normal discourse, these potential forms are relatively uncommon perhaps in order to avoid such ambiguities (because the lexical forms are recognized first in our mental parsing before syntactical processing), and for this purpose another potential suffix "koto-ga-deki" (i.e., ORkotogadeki, NARAB-kotogadeki, respectively)[35] , which will be discussed in the next paragraph, is much more frequently used.

There are two alternative syntactic constructions for potential. The first, "koto-ga-deki" (50a), is in some way similar to the English sentence pattern of "it is possible that ...," semantically and syntactically. This "koto-ga-deki" is further analyzable such that "koto" is a formal noun like English "it," "ga" is a nominative postposition, and "deki" is a suppletive "potential" form of Sahen-verb "s(uru)." However, this sequential expression appears more idiomatic as an inseparable single unit. Consequently, there are two ways to treat this "koto-ga-deki." One is to segment it into three discrete elements, and the other is to treat it as a single indivisible unit that is of the same level as other "words." Though we adopt the first option, both treatments are equally qualified from our point of view in order to carry out our retrieval experiments, because no case change is involved in the original main phrase. Thus, we simply put a secondary indexing marker for syntactical potential @POTN at the place of the segmented element of "deki."

(50) a. John-ga Sono Eda-o OR-ϕ-ru <u>koto-ga-deki</u>-ta.
John-NOM the branch-ACC break-TRANS-Adj FmN-NOM-POTN-PAST
('John could break the branch.')

[35] With negation, "-e" suffix is more commonly used as follows:

g'. John-wa Sono Eda-{o, ga} OR-ϕ-<u>e</u>-naka-ta. (ka-ta=>katta)
John-TOP the branch-ACC break-TRANS-POTN-NEG-PAST
('John was not able to break the branch.')

i'. Kodomotati-ga Umaku Itiretuni NARAB-ϕ-e-naka-ta. (ka-ta=>katta)
kids-NOM well straightly line.up-INTR-POTN-NEG-PAST
('The kids could not line up very well.')

However, we can disambiguate these cases by designating the object item with proper case marking. Furthermore, in type E (g'), the object more frequently receives "-ga" postposition, so the standard operation for transitivization can be applied for paraphrasing with no harmful effects.

b. John-ga Sono Eda-o OR-ϕ-i E-ta.
John-NOM the branch-ACC break-TRANS-Adv POTN-PAST
('John could break the branch.')

The other potential form, V(adverbial)+"e" (50b), is a rather uncommon expression, and sounds like a literal translation from a Chinese text. As described in footnote 20 in Section II.d, this "e" is attached to the adverbial form of the verb, i.e., at the point in time when all derivational and inflectional suffixation have been completed, so that this "e" behaves as if it is a separable independent verb that composes a V+V type compound by following a preceding verb. "E" is homonymous with a free verb "e," 'get,' and the same kanji character (得) is assigned to both cases. We may call this kind of potential "compounding potential." Thus, in the same way we would we treat a standard verbal compound, we segment "e" as a distinct element, and as in the case of "koto-ga-deki," we put a secondary indexing marker @POTN at the place of this "e" potential.

To summarize, we assign three distinct syntactical markers for verbs, i.e., "@CAUS-l" for causative, "@PASS-l" for passive, and "@POTN" for potential. (Note that the potential marker, @POTN was also used for verbal nouns as described in the next subsection.)

d. Indexing Markers for Verbal Nouns

In section II.G, we discussed the grammatical relations in a verbal noun construction. A verbal noun has both nominal and verbal characteristics. As a verbal element, the verbal noun has its own argument structure, and this is observed when it appears either as a predicate or as a head of a genitive noun phrase. Thus, we focus on these two grammatical relations in implementing the verbal noun query formulation. Let us look at each case.

When a verbal noun is used as the predicate of a sentence, it appears incorporated with a copula (Sahen-verb). Depending on whether a verbal noun or a verb is used for the query, the retrieval performance will exhibit a difference for two reasons. First, in Japanese verbal nouns are lexically less ambiguous than verbs. A verbal noun with a simple meaning may have a corresponding verb, but verbal nouns tend to express more complex and contextually specific meaning. Therefore, in such cases, a corresponding verb may not exist. For example, KAITEN-sase 'rotate (caus.)' in (51b) may be paraphrased by a native verb "MAWA-s" 'rotate (tr.),' but there is no single verb which corresponds to "TAIDEN-s" 'be-electrically-charged," so the only way to paraphrase this with a native verb is to use a phrase like "Denki-o OBI-ru" (electricity-ACC be.charged). Second, verbal nouns are preferably used in a style of written text like the newspaper article in our experiments. Therefore, semantic specificity and frequent use in text will result in a higher effectiveness of query performance for verbal nouns than for verbs, even when they display a similar meaning. However, more importantly, it is not yet clear whether valency control query strategies more effectively improve retrieval for verbal nouns than for verbs (or vice versa). Such improvement depends on how frequently the causative/passive/potential is applied to the sentences which are significantly related to the relevant topic of the text.

Causativization, passivization, and potentialization for verbal nouns are realized by syntactic means by changing the form of the copula (i.e., Sahen-verb). Thus, when a Sahen-verb "s" becomes "s-ase," "s-are," or "deki," it represents the causative, passive or potential, respectively. These are exemplified in (51a,b,c). The indexing program finds the incorporation of verbal noun with Sahen-verb, and puts a secondary indexing marker @PREDICATE_VN at the place of Sahen-verb, and if the Sahen-verb shows the above causative, passive, or potential form, the program places the marker @CAUS-w, @PASS-w, or @POTN, respectively. Thus, the syntactic potential marker @POTN is used for both native verbs (e.g., "koto-ga-deki") and for verbal nouns ("deki").

(51) a. Atarasii Suisei-ga HAKKEN-sare-ta.
new comet-NOM discovery-PASS-PAST
('A new comet was discovered.')

b. John-ga Mootaa-o KAITEN-sase-ta.
John-NOM motor-ACC rotating-CAUS-PAST
('John rotated the motor.')

c. John-wa Bouenkyou-de sono Suisei-o HAKKEN-deki-ta.
John-TOP telescope-INSTR the comet-ACC discovery-POTN-PAST
('John was able to discover the comet by a telescope.')

Next, let us consider the case of the genitive noun phrase. When a verbal noun is the head of a genitive noun phrase, it appears as the right-most element because Japanese follows the left-branching (i.e., right-headed) rule. (52a,b) shows the examples.

(52) a. Suisei-no HAKKEN
comet-GEN discovery
('discovery of the comet')

b. Einstein-no HAKKEN
Einstein-GEN discovery
('Einstein's discovery')

The *no*-phrase is often interpreted as the subject (52b) or the object (52a) of the head verbal noun.[36] Therefore, there is a problem of grammatical ambiguity, such that we may misidentify the subject as the object (or vice versa)[37] when the verbal noun is transitive. For example, the subject, e.g., 'Einstein' in (46b), might be misjudged as the object of the verbal noun ('discovery'). Table 10 shows the frequency data of 100 samples of genitive noun phrases, showing the frequencies of grammatical relations.

Table 10. Frequencies of grammatical relations in genitive NPs.

Intransitive Subject	21
Transitive Subject	17
Transitive Object	54
Ambiguous (Intrans.Subj. / Trans.Obj.)	8
(Total	100)

From this sample data, we obtain the probability 21.5% (17/(17+54+8)) of misjudgment of a transitive subject when we uniformly classify it as an

[36] This is not always the case. In a sampling of 164 genitive noun phrases, 48 cases (i.e., 29%) did not represent argument relationships like those in (1) (cf. (46)):

(1) Rosanzerusu-no HOUKI (Los.Angeles-GEN riot) 'riot of Los Angeles'

However, this kind of noun phrases will not cause much harm for retrieval because they express other kinds of contextual relations (e.g., in (1), locative) which are not the target of the valency control query strategy. Furthermore, in many cases, the original case-postposition is explicitly attached to the genitive postposition to make the relationship clear, as seen in (1'):

(1') Rosanzerusu-de-no HOUKI (Los.Angeles-LOC-GEN riot) 'riot in Los Angeles'

This formation is used also for indirect objects, as in (2):

(2) Medaru-no Kagakusya-e-no ZOUTEI (medal-GEN scientist-DAT-GEN-presentation) 'presentation of the medal to the scientist'

Our indexing program ignored the genitive postposition in such combined postpositions.

[37] This situation contrasts with Roeper & Siegel's (1978) First Sister Principle for verbal compounds - only arguments at the first-sister position (i.e., objects if the verb is transitive) can be attached to their head elements. Kageyama (1993, pp.217-222) argued that this constraint is not only applicable to lexical compounds, but also to syntactical compounds as seen in (1) and (2) (The notation of syntactical compounds follows Kageyama's.) :

(1) [Suisei:HAKKEN]-no Nyuusu ('news of discovery of the comet')

(2) *[Einstein:Hakken]-no Nyuusu ('news of Einstein's discovery').

object. This is the method we followed in our experiments - when the verbal noun is transitive (including the ambiguous case), the @GEN_ACC marker is indexed, otherwise the @GEN_NOM is marked. Although the significant majority of cases correctly occur on the transitive object side, still this error rate is not negligible, and might degrade the effect of the query strategy on verbal nouns.[38] (Also see Subsection D.5 about the query formulation for VNs.) Note that 8% of verbal nouns polysemously overlap in the intransitive and transitive. In such cases, there is absolutely no way to determine the category because of the lack of a morphological device to distinguish the two cases. This is a quite a different situation from verbal doublet making. For example, as a paraphrase of (53a), both (53b) and (53c) are equally qualified.

(53) a. Boueki-no KAKUDAI
trade-GEN increasing
('increase of trade')

b Boueki-ga KAKUDAI-suru.
trade-NOM increase(intr.)-PRES
('Trade increases.')

c. (X-ga) Boueki-o KAKUDAI-suru.
(X-NOM) trade-ACC increase(tr.)-PRES
('X increases trade.')

To summarize, we assign six secondary indexing markers for verbal nouns, i.e., "@CAUS-w" for causative, "@PASS-w" for passive, "@POTN"

[38] This negative effect may be serious if the baseline query has a genitive construction, and the query strategy is applied to derive a predicative form. Conversely, in our experimental setting, baselines are set as plain predicative sentences for across all strategies, and genitive strategy derives genitive strategic patterns from the baselines. Therefore, in this case, above ambiguity problem can be absorbed and avoided, however such a strategy has a weak grammatical specificity. Indeed, our experiments showed that genitive strategies on verbal nouns did not work so effectively as other valency strategies on verbs or verbal nouns do. (See IV.B.5.)

(this was also used for verbs) for potential, and @GEN_NOM for genitive with nominal, @GEN_ACC for genitive for accusative, and @PREDICATE_VN for predicative verbal nouns.

7. Summary of Text Indexing

This section described the range of aspects of text indexing in our experimentation. In the first four subsections, we described four basic tasks and their techniques, namely sentence identification, word identification, morphological analysis, and syntactic analysis. After sentences and individual words in a text are identified, morphological analysis identifies subparts like affixes, then syntactic analysis determines the structural relationship between the words in each sentence.

The two subsections that followed, B.5 and B.6, described the index terms identified by the techniques described above. Subsection B.5 described the primary indexing terms, i.e., keywords, that are extracted from the text. Throughout subsection B.6, we presented various secondary indexing methods based on the grammatical features or relationships that characterize Japanese verbs and verbal nouns. Since these features and relationships are indispensable elements for the grammatical paraphrasing specifications of valency control, let us extensively review these secondary indexing units. Table 11 shows the list of secondary indexing markers used in our experiments.

Table 11. List of secondary indexing markers.

Indexing Marks	Meaning
Grammatical Relationship	
@SUBJ	subject
@DO	(direct) object
Lexical Class	
@TRANS	unmarked transitive, or transitivized by "-e"
@INTRANS	unmarked intransitive, or intransitivized by "-e"
@CAUS-s	short causative
@PASS-s	short passive
Syntactical Class	
@CAUS-l	(long) causative (for V)
@PASS-l	(long) passive (for V)
@POTN	potential (for V and VN)
Verbal Noun Class	
@PREDICATE_VN	predicative verbal noun by incorporation
@CAUS-w	causative of VN copula
@PASS-w	passive of VN copula
@GEN_NOM	nominative argument of genitive VN(intr.) phrase
@GEN_ACC	accusative argument of genitive VN(tr.) phrase

Looking over these characteristics, all markers in Table 11 are practically identifiable by morphological analysis, although grammatical ambiguity problems might result in certain misjudgments, depending on the target of the analysis. When a verb or a verbal noun is used as the head of a predicate, the subject (@SUBJ) and the object (@DO) (when the verb is transitive) are normally expected to be present in the sentence, if not omitted structurally (e.g., in an interrogative) or by discourse. (We did not use indirect objects in our experiments because the number of ditransitive verbs are very limited - indeed our test queries did not include applicable ditransitive verbs.)

Next, we focused on the verbal functions of valence changing. These functions are divided into two classes, i.e., lexical and syntactic. The lexical class of secondary markers are applied to verbs and designate their transitivity, i.e., transitive or intransitive. The short causative is an amalgam of syntactic and lexical representation, though the lexical short causative is empirically dominant in all short causative formations. We decided to index the short causative separately (@CAUS-s) from other lexical transitive formations, for possible separate use. Although the short passive does not have such grammatical ambiguity, it behaves as an absolute intransitivizer, just as the short causative works as an absolute transitivizer. Thus, we also index the short passive separately (@PASS-s). In contrast, the "short potential" "-e" morpheme does not show such absolute characteristics and can appear as both transitive and intransitive, depending on the root. We grouped together the unmarked transitive and verbs transitivized by "-e" (@TRANS), and did the same for the unmarked intransitive and verbs intransitivized by "-e" (@INTRANS).

The syntactic secondary marker group consists of three markers: (long) causative (@CAUS-l), (long) passive (@PASS-l), and potential (@POTN). These formatives are identified in the segmentation process in order to separate every word by morphological rules. Although the long passive construction may signify either passive or potential, the majority of instances have a passive meaning. The potential marker is used not only with verbs, but also with the Sa-hen verb, which is used in conjunction with verbal nouns.

The last class of secondary markers was designed especially for verbal nouns. Since the verbal noun has a morphologically invariant nominal form, its (in)transitivity is not morphologically distinguishable. All the secondary markers for verbal nouns have syntactic characteristics. When a verbal noun is incorporated with a copulative verb "s" it works as a predicate head, and a mark @PREDICATE_VN is indexed at the position. When a verbal noun is not predicative but is the head of genitive noun phrase, the modifier's grammatical relation to the head is predicted by the transitivity of the verbal noun head. If it is predicted as the subject of the intransitive verbal noun, or the object of the transitive verbal noun, @GEN_NOM or @GEN_ACC is marked, respectively.

C. Characteristics of Text Collection

To characterize the text collection used in our experiments, we should describe both its quantitative characteristics (e.g., subject area) and its quantitative characteristics (e.g., collection size). As an ideal, the collection should be qualitatively general, and quantitatively large enough to prove its practicality in real applications. The size of our text collection is reasonably large. Though it is not a collection of "heterogeneous" sources, it nevertheless shows similar general qualitative characteristics to some degree. Our test collection, which was also used in the TIPSTER project (Harman, 1992), consists of newspaper articles from the Nikkei Shinbun (1991 1/1 -- 1991 12/20). The subject area is general, but more or less with an emphasis on the business domain (comparable to the Wall Street Journal in English). It contains a total of 151,650 documents (178MB). Table 12 shows the frequencies (per document, and per sentence) of words, sentences, major lexical categories (i.e., parts-of-speech and their subcategories) such as nouns, (transitive and intransitive) verbs, verbal nouns, and adjectival nouns, and several key morpho-syntactical categories such as causative (CAUS), passive

(PASS), and potential (POTN), in our text collection. Let us describe lexical and syntactic categories, separately.

Table 12. Average frequencies of major grammatical categories.
(Documents with no sentences such as tables and lists are treated as "missing" cases. These counts 21,573 documents (14.2%) of total 151,650 documents. Thus, the actual number of documents are 130,077. Sahen-verbs are excluded from counts of verbs. Symbols such as commas and periods are not counted as words.)

category	average per doc (S.D.)	per sentence (/S)	% in total words
<Lexical Category>			
N (nouns)	100.0 (81.5)	9.17 N/S	36.3
AN (adj.nouns)	8.1 (9.2)	0.74 AN/S	2.9
VN (verbal nouns)	29.3 (26.1)	2.69 VN/S	10.6
VNpred (predicative VN)	7.1 (6.2)	0.65VNpred/S	2.6
V (verbs)	17.2 (18.6)	1.58 V/S	6.3
Vt (trans. verbs)	10.4 (11.2)	0.95 Vt/S	3.8
Vi (intrans. verbs)	6.8 (8.1)	0.62 Vi/S	2.5
(subtotal)	(154.6)	(14.18)	(56.2)
W (words)	275.2 (230.1)	25.25 W/S	--
S (sentences)	10.9 (9.7)	--	--
<Syntactical Category>			
CAUS-l (causive(V))	0.083 (0.33)	0.8% CAUS-l/S	--
CAUS-w (causative(VN))	0.169 (0.50)	1.6% CAUS-w/S	--
PASS-l (passive(V))	0.870 (1.53)	8.0% PASS-l/S	--
=POTN-l (long potential (='(r)are'))			
PASS-w (passive(VN))	0.502 (1.09)	4.6% PASS-w/S	--
POTN (except POTN-l)	1.229 (1.94)	11.3% POTN/S	--
(subtotal)	(2.853)	(26.2%)	

Regarding lexical categories, we notice the following characteristics. First, the per-sentence values of verbs and verbal nouns are larger than 1, probably because they can be used in a verbal compound, a compound (i.e., paralleled) sentence, or a complex (i.e., embedded) sentence. Assuming several nouns are assigned to the subject and the objects in a sentence, the value of 9.2 nouns per sentence implies that probably half to two third of nouns are used in adjunct phrases. Concerning the percentage of categories of words used, 56% of words in a document fall into the above major categories. 53% (=17.2/(17.2+7.1+8.1)) of predicative elements (i.e., V, VN(predicative), and AN) fall into the V (i.e., native verb) category. Note that the value of AN (=8.1) includes the number of its nominal uses, so this estimated rate has a small negative bias. Since our valency control strategies mainly focus on verbs, this relatively high usage of verbs suggests the practical applicability and potential retrieval effectiveness (because if the usage of verbs is too low, no verb strategy can contribute to the performance improvement) of these strategies.[39] The strategic applicability of queries will be discussed in the next section, and actual performance measurements of experiments will be demonstrated in the next chapter.

Concerning verb subcategories, i.e., transitive and intransitive, we should keep in mind that the data includes not only doublet verbs, but also non-doublet verbs (except Sahen verbs). Therefore, the frequency difference between transitive (Vt) and intransitive (Vi) is not a simple indicator of the choice in possible transitivity alternatives, though the majority of verbs in Japanese are doublet type. (See also Table 2 in Chapter II.) Frequency values of

[39] It is an interesting question how a typological configuration of the part-of-speech system affects the valency control query formulation and the performance, however we do not explore this topic in this dissertation.

transitive and intransitive verbs are certainly reasonable - around one transitive verb in a sentence (0.95 Vt/S), and the intransitive value is around two third of the transitive value (0.62 Vi/S).[40]

Regarding syntactic categories, causative is only the case (even if we add their verb and verbal noun cases) which occurred less than once in a document, and both the passive case (if we add the verb and verbal noun cases), and the potential case exceeded 1 per document. Since the potentialization of a vowel stem verb is identical to the one of long passive (i.e., -rare), their classifications (@PASS-l and @POTN-l, respectively) eventually overlap. However, the great majority (95%, including middle, as we already described in the previous section (III.6.c.)) can be treated as passive rather than potential, and therefore the statistical figure in this table will not be significantly affected even if we make this adjustment.

Next, let us observe per sentence values of syntactic categories. In Table 12, we represented these numbers in percent (%) because they are eventually equivalent to the probability of their occurrences when we randomly pick up a sentence in the corpus. For example, if we look at a sentence, there is an 8.0% of chance of being passive (i.e., PASS-l/S for verbs). The data of causative and passive exhibit a great contrast to the data of transitive and intransitive verbs, in two ways. First, the value difference in the causative-passive pair is strikingly large (i.e., passive is around 10 times more frequent than causative (8.0% vs. 0.8%)), but the value gap between transitive and intransitive is relatively small (i.e., frequency of intransitive is two third of the data of transitive (0.62 vs. 0.95)). Second, the frequency inequality

[40] In Svartvik's (1966, p. 62) data, 60% of English active clauses are transitive (or ditransitive), and 23% are intransitive. In the same data table, he reported that 12 % of English clauses are passive, and 88% are active.

relations between causative and passive, and between transitive and intransitive are opposite in terms of their valencies. In other words, in lexicon, bivalent patterns (i.e., transitive) are more frequent than monovalent patterns (i.e., intransitive), however, in syntax, bivalent patterns (i.e., causative) are less frequent than monovalent patterns (i.e., passive). It is interesting to see how these characteristics relate to the performance of valency control retrieval strategies. We will see such experiments in the next chapter.

D. Query Formulation

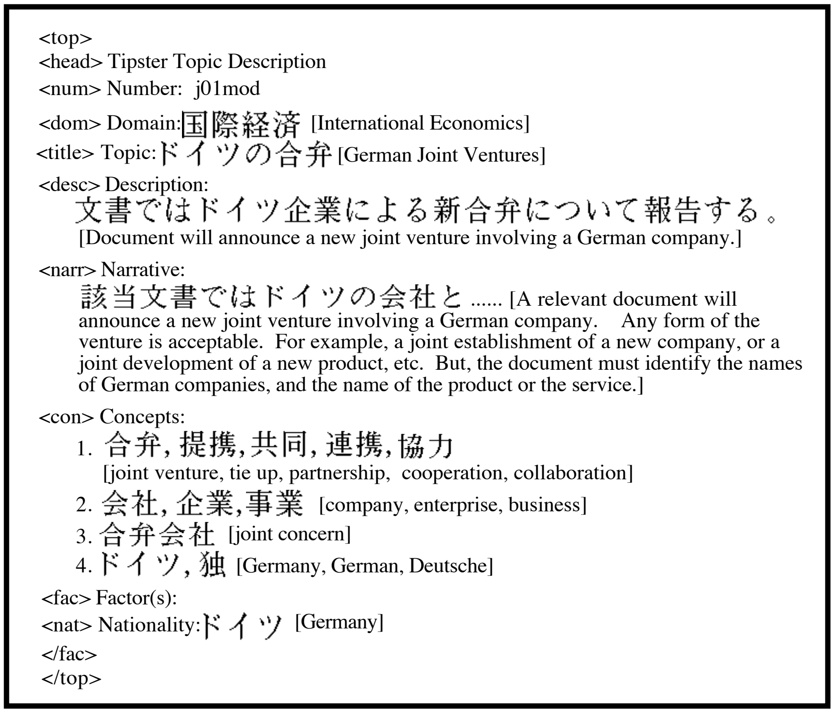
The retrieval process applied on an already indexed text database is divided into three stages. First, the user's *information need* must be expressed in a certain way. Second, the expressed information need must be translated into a certain query form, which must be an acceptable form for the target retrieval system. This query expression may be augmented by certain query formulation strategies that we may wish to develop and prove the effectiveness of. From the point of view of retrieval experimentation, two distinct sets of queries must be provided: a baseline query and a target query. The performance of the target query should be compared with the baseline performance in order to measure relative performance improvement (or degradation). Finally, these queries are entered into the retrieval system to run on the database. In this section, we describe the first two stages, i.e., expressing an information need and translating the information need into a query formula, as preparatory steps before the query is run.

1. Representation of Information Need

An information need can be expressed in various ways. Two typical ways are as a list of keywords and as a natural language description. A keyword list is a collection of concepts which are expressed by semantically relevant words or phrases. On the other hand, a natural language description is written down as a description in plain language, say, in English or Japanese. We adopted an integrated representation of these two methods, which was developed in the TIPSTER project (Harman, 1992). An example of a TIPSTER-type information description (called a *topic*) is shown in Figure 2.

A TIPSTER-type topic is structured by various fields, and each field is labeled by a tag. In Figure 2, keyword-style information is expressed in the "<con>" (concept) field. Keywords are conceptually clustered in several groups. Thus, each group consists of synonymous words. On the other hand, natural language type information is expressed in three detail levels. The shortest description is the title, "<title>," which is typically a single noun phrase. The middle level is the description, "<desc>," which usually consists of one or a few sentences. The most detail level is the narrative, "<narr>," which usually contains several sentences that often describe the specific conditions used to judge the relevance of a retrieved document. Note that keyword-style specification and natural language description may be regarded as different representations of the same information need. Keywords in the <con> field are supposed to cover the total information need. Therefore, in our experiments, our query construction operation used keywords in the <con> field as a major source of query terms, and the natural language descriptions in <title>, <desc>, and <narr> are used as supplemental sources of keywords. Other categorical fields such as "<dom>" (domain), "<fac>" (factor), and "<nat>" (nationality) were not used in our experiments because their purposes are too specialized and difficult to be incorporated with other general fields.

We had a total of 16 Japanese topics which were developed in the TIPSTER project. We had to select a subset of topics based on strategy-applicability, as we will describe in following discussions. These selected TIPSTER topics are translated into query forms that are acceptable for an experiment platform, INQUERY retrieval system.



```
<top>
<head> Tipster Topic Description
<num> Number:  j01mod
<dom> Domain:国際経済 [International Economics]
<title> Topic:ドイツの合弁 [German Joint Ventures]
<desc> Description:
    文書ではドイツ企業による新合弁について報告する。
    [Document will announce a new joint venture involving a German company.]
<narr> Narrative:
    該当文書ではドイツの会社と ...... [A relevant document will
    announce a new joint venture involving a German company.   Any form of the
    venture is acceptable.  For example, a joint establishment of a new company, or a
    joint development of a new product, etc.  But, the document must identify the names
    of German companies, and the name of the product or the service.]
<con> Concepts:
    1. 合弁, 提携,共同, 連携,協力
       [joint venture, tie up, partnership,  cooperation, collaboration]
    2. 会社, 企業,事業 [company, enterprise, business]
    3. 合弁会社 [joint concern]
    4. ドイツ, 独 [Germany, German, Deutsche]
<fac> Factor(s):
<nat> Nationality:ドイツ [Germany]
</fac>
</top>
```

Figure 2. Structure of Japanese TIPSTER topic.

2. Retrieval Operators and the Three-Level Structure of Queries

Once an information need is described in a TIPSTER-style topic, we can transform the topic into a query form. The query form must be acceptable by the target retrieval system - in our case, the INQUERY system (Turtle, 1991; Callan, Croft, and Harding, 1992). An important point to realize is that in order for post-coordination to achieve the full-strength of the retrieval strategy, the retrieval system must be well-equipped with operational devices that encode the strategy into a query form. Traditionally, such functions are implemented as *retrieval operators* for a specific retrieval system. The basic characteristics of operators are determined based on the "retrieval model," which is the theoretical basis of the retrieval operation. For example, the Boolean retrieval model has typical Boolean operators such as AND, OR and NOT. The vector-space model and the simple probabilistic model typically do not have operators, because these model assume that the set of given terms as a whole is the ultimate input to the system for computing a single degree of relevancy for each document. In other words, they heavily rely on pre-coordination to improve retrieval performance. On the other hand, INQUERY's *inference network model,* which is a probabilistic Bayesian model for document retrieval, provides various query operators to define the relationship between two or more terms (i.e., network nodes). INQUERY's operator capability facilitates implementing various query formations in our experiments.

The form of query in INQUERY is called "structured query" because each query is organized by a number of query terms and operators which are applied to the terms. The general form of query syntax is as follows in (54):

(54) a. QUERY := OP(QUERY_1, QUERY_2, ..., QUERY_n) | term

b. OP := #sum | #passageN | #proxN | #lit | others [N is a number.]

Here, OP stands for a retrieval operator, and "term" is either a primary indexing term, i.e., a keyword, or a secondary indexing term such as @SUBJ. Inside the operator OP, n subqueries can be placed as the arguments in nested manner. The number of arguments, i.e., n, differs depending on the kind of operator, and it is usually variable. INQUERY provides the various operators listed in (54b) (limited to only those used in our experiments). The operator "#sum" returns the mean of argument "beliefs" which is attached at each node of the inference network. The operator "#lit" means a literal, so it has no effect on the arguments. However, an argument may contain special symbols like "@," as in our secondary indexing markers. The operator "#passageN" looks for a pattern of arguments which are arranged in a window of size N. The operator "#proxN" (or simply #N) stands for the proximity relationship, which means that each adjacent argument pair must be at a distance N or less, and occur in the order specified. Thus, we have two ways to specify a pattern of term sequences. Since the arguments of the passage operator #passageN can be arranged freely - the distance between two given terms can be elastic within the limit of the window's boundary - it is useful to specify a grammatical pattern such as a subject-object relationship in a sentence, because we do not know how many extraneous elements occupy positions between the subject and object. On the other hand, the proximity operator #proxN restricts the size of every gap between argument pairs. Therefore, it is good for specifying more rigid patterns like compounds or idioms or proper names. Note that these two windowing operators work as purely mechanical string manipulations, and are not implemented by any direct linguistic representation to drive the operational mechanism of retrieval. Thus, all linguistic relationships must be communicated externally to the retrieval mechanism. In section B.1 of this chapter, we already described how a series of protecting fillers is inserted between sentences to prevent a matching pattern from overflowing the sentence boundary. Another example of this kind of limitation is that we do not have an effective way to guarantee that two elements with subject and object markers grammatically correspond within the same clause, so as not to be split over the main and subordinate clauses. We point out that such a weak pattern matching capability occurs not only in our experimental system, but also in most traditional text retrieval systems. However, as this is an issue of implementation, we do not explore it in this dissertation.

A TIPSTER-style topic is transformed into an INQUERY query construction, which has a three-level query structure containing the following three subqueries: 1) *keyword-level baseline subquery*, 2) *sentence-level baseline subquery*, and 3) *strategic subquery*. Figure 3 is the outline of this three-level query structure using operator notation. The keyword-level subquery is basically a collection of keywords that are conceptual constituents of the topic. Its performance serves as the most basic level of the baseline, i.e., the *keyword-level baseline*. This means that no active information is involved in any utilization of grammatical knowledge at this baseline level.

Sentence-level subquery is provided to compensate for the bias of sentences with no grammatical specifications. Thus, it specifies only a pair of a noun and a verb, but does not contain any secondary indexing markers used for designating a grammatical relationship. Combinations of a verb (or a verbal noun) and its potential argument (i.e., subject or object) are coupled inside a #passageN operator. In Japanese, in the standard word order a verb always follows an argument, and always comes at the sentence (or clause) end.

The strategic subquery is that subquery which implements the query strategy. Essentially, this subquery is an arrangement of a verb, the argument noun, and secondary indexing markers that establish and indicate the grammatical relationship of the verb and the noun.

Thus, these elements are the same set as those used in the sentence-level baseline subquery, except for the grammatical markers. The effect of a strategy encoded in this subquery should be compared with the performance of the sentence-level subquery, if we want to measure the genuine effect of a valency control strategy, which should not be biased by the recognition of the sentence itself. However, comparison with the keyword-level baseline also has a practical meaning in that it shows the boost of effectiveness from the most basic keyword-based query to the full-fledged grammatical query. In the following subsections we discuss the method for making each subquery.

```
#sum(
  kw1,  kw2,  . . . ,   kwn,          }  <======= Keyword-level Subquery
  #passage(N1, V1), #passage(N2, V1), ..., #passage(Np, V1), \
          ..., #passage(Ni, Vj), ...                          }  <= Sentence-level
  #passage(N1, Vq), #passage(N2, Vq), ..., #passage(Np, Vq), /      Subquery

  #passage( "representation of valency control strategy" ),  \
                 . . . . .                                    }  <= Strategic
  #passage( "representation of valency control strategy" )   /      Subquery
  )
```

Figure 3. Three-level query structure.

3. Construction of Keyword-level Baseline Subqueries

The keyword-level baseline subquery is the most basic part of the query, and consists of a set of keywords that give the "keyword-level baseline" performance. The most important characteristic of this subquery is that it includes all verb keywords and all their argument nouns which are entities query strategies are exclusively applied. The purpose of the rest of keywords is just to describe the other portion of information need, and they do not involve in the effect measurement of query strategies. The transformation of a TIPSTER-type topic into a keyword-level subquery is a manual process, but follows three simple guidelines:

(i) Select pairs of verbs (or verbal nouns) and their semantically associable arguments (no matter whether the pair exists in texts, or not) from the <con> field. For verbal nouns, semantically corresponding verbs also should be tried out, and vice versa. Since verbs and verbal nouns have synonymous correspondences, but are grammatically distinct entities to be examined in different series of experiments, their collection must be treated separately.

(ii) Using a general thesaurus, reinforce already selected query terms by synonymous (including "co-hyponymous") relationships. Note that this synonym expansion contributes to recall, but its aim differs from that of our grammatical query strategies, which must be regarded as precision enhancement devices. The purpose of this expansion is simply to increase the chance of strategic applicability.

(iii) Add other significant keywords, if any exist and were not included before, from the <title>, <desc>, <narr>, and <con> fields.

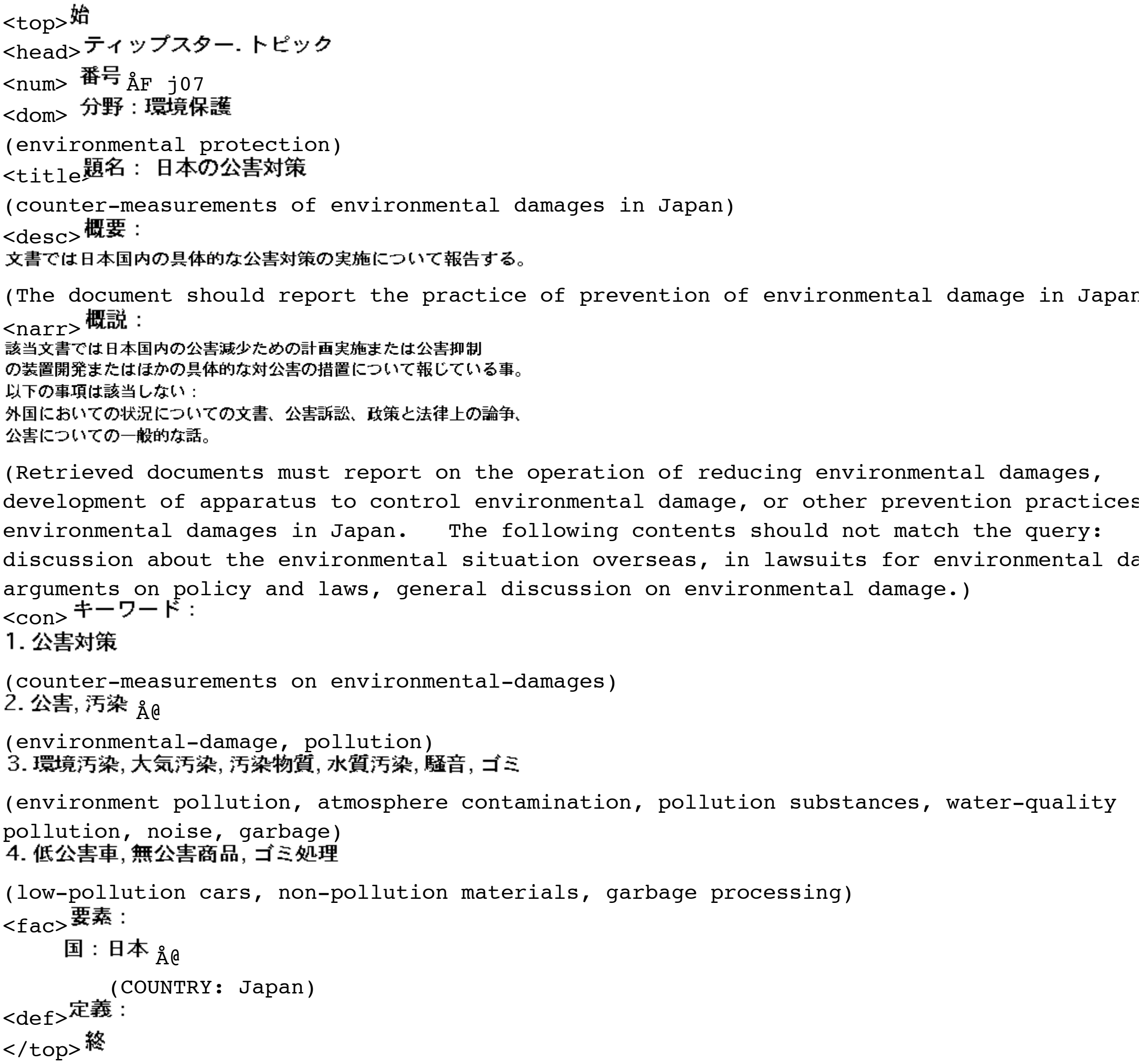
We will show a real application of the above guideline later in this subsection.

As we see in these guidelines, selecting keywords from a topic inevitably involves assessing the term's strategic applicability, and we eventually evaluate the topic's applicability to the strategy. Therefore, we derive the following two conditions in order to distinguish acceptable topics from ones to be rejected:

a. existence of verb-noun direct products: This is a requirement of applicability of the valency control query strategy. This condition implies that both verbal elements (verbs and verbal nouns) and their argument nouns should be non-empty. Without these elements, no valency strategy can be applied to the query.

b. existence of corresponding native verbs: In many cases, verbal concepts are given as verbal nouns. In such case, corresponding native verbs must exist for assessment to be possible, since our main objective in experiments is to measure the effect of the valency changes of verbs.

By applying above criteria, we selected 6 applicable topics from a total of 16 TIPSTER topics. Once a topic is selected as strategically applicable, we can transform it into a query representation. To make the following discussion concrete, let us examine an actual example of an acceptable topic, j07 (Figure 4).



```
<top>始
<head>ティップスター. トピック
<num> 番号ÅF j07
<dom> 分野：環境保護
(environmental protection)
<title>題名： 日本の公害対策
(counter-measurements of environmental damages in Japan)
<desc>概要：
文書では日本国内の具体的な公害対策の実施について報告する。
(The document should report the practice of prevention of environmental damage in Japar
<narr>概説：
該当文書では日本国内の公害減少ための計画実施または公害抑制
の装置開発またはほかの具体的な対公害の措置について報じている事。
以下の事項は該当しない：
外国においての状況についての文書、公害訴訟、政策と法律上の論争、
公害についての一般的な話。
(Retrieved documents must report on the operation of reducing environmental damages,
development of apparatus to control environmental damage, or other prevention practices
environmental damages in Japan.   The following contents should not match the query:
discussion about the environmental situation overseas, in lawsuits for environmental da
arguments on policy and laws, general discussion on environmental damage.)
<con>キーワード：
1. 公害対策
(counter-measurements on environmental-damages)
2. 公害, 汚染 Å@
(environmental-damage, pollution)
3. 環境汚染, 大気汚染, 汚染物質, 水質汚染, 騒音, ゴミ
(environment pollution, atmosphere contamination, pollution substances, water-quality
pollution, noise, garbage)
4. 低公害車, 無公害商品, ゴミ処理
(low-pollution cars, non-pollution materials, garbage processing)
<fac>要素：
     国：日本 Å@
        (COUNTRY: Japan)
<def>定義：
</top>終
```

Figure 4. A sample topic.

The list in (55) shows direct products of verbal elements and their argument nouns, which were extracted from the topic j07 in Figure 4. Formula (55a) represents the original verb-noun combinations from the topic as extracted by guideline (i). Terms in (55a) are expanded by guideline (ii), and the result is shown in (55b) in verbal domain, and in (55c) in verbal noun domain. Terms in (55b) (or (55c)) are listed in the keyword-level baseline subquery. Furthermore, they represent the core of the generated query, and will be used in the production of sentence-level and strategic subqueries.

(55)

a. N×V ={{環境, 大気, 水質}×{汚染}, {ゴミ}×{処理}}
{ {environment, atmosphere, water quality}×{pollution}, {garbage}×{processing} }

b. N×V = {{大気, 空気, 水質|水, 河川, 湖水, 海洋}×{汚す/汚れる},
{ {atmosphere, air, water quality, river, lake, ocean}×{contaminate/be.contaminated},
{ゴミ, 廃棄物}×{棄てる, 捨てる, 焼く/焼ける}}
{garbage}×{dump, dump, burn(tr)/burn(intr)} }

c. N×V = {{大気, 空気, 水質, 水, 河川, 湖水, 海洋}×{汚染},
{ {atmosphere, air, water quality, river, lake, ocean}×{contamination}
{ゴミ, 廃棄物}×{廃棄, 焼却}}
{garbage, waste}×{dumping, burning} }

Finally, to complete the keyword-level subquery for this sample topic j07, we add the word "対策" 'counter-measurement' into the subquery using guideline (iii). Although this keyword is not a part of the verb-argument pairs, it appears as a significant keyword in the <title> field. After collecting these keywords (i.e., no more direct product notation), the final form of the keyword-level subquery for native verbs is as follows in (56):

(56) {大気, 空気, 水質, 水, 河川, 湖水, 海洋, 汚す_@ORG,
ゴミ, #1(廃棄 物), 棄てる, 捨てる, 焼く_@ORG , 対策 }

In (56), we notice two technical points. First, the "@ORG" marker, which signifies the original (i.e., not "normalized" in transitive alternations) verb form (see the discussion in section B.3 in this chapter), is attached to the verbs, e.g., 汚す_@ORG. We used this verb coding because we did not want to introduce any grammatical effect on the keyword-level baseline[41]. The

[41] When the verb has a doublet partner, we may attach the @ORG marker to either the transitive or intransitive form. Because of this reason, slight performance fluctuation may occur. We chose the case to make the verb and corresponding verbal noun be semantically consistent.

second point is that the proximity operator "#1" was applied to "廃棄+物" ('to waste'+'thing'='garbage') because this is a compound, and its two elements are tightly coupled.

4. <u>Construction of Sentence-level Baseline Subqueries</u>

In the process of keyword selection described in the previous subsection, we eventually produced a set of direct products of verbs and their arguments. So, basically speaking, all that is required to construct sentence-level subqueries is simply to apply the #passageN operator to each combination. An example of sentence-level subquery (a set of "passages") corresponding to (55b) is shown in (57):

(57)

```
#passage14(大気, 汚す_@ORG),
#passage14(空気, 汚す_@ORG),
#passage14(水質, 汚す_@ORG),
#passage14(水, 汚す_@ORG),
#passage14(河川, 汚す_@ORG),
#passage14(湖水, 汚す_@ORG),
#passage14(海洋, 汚す_@ORG),
#passage14(ゴミ, 棄てる_@ORG),
#passage14(ゴミ, 捨てる_@ORG),
#passage14(ゴミ, 焼く_@ORG),
#passage14(#1(廃棄, 物), 棄てる_@ORG),
#passage14(#1(廃棄, 物), 捨てる_@ORG),
#passage14(#1(廃棄, 物), 焼く_@ORG)
```

Here, the passage operator #passageN was applied to the verb-noun combinations. The difference in the linguistic characteristics of sentences and compounds reflects their use of operators, i.e., #passageN and #proxN, respectively.

Note that some of these expanded verb-noun combinations might not significantly contribute to retrieval performance. We developed two effective approaches to optimize the query by enhancing only positive passages, namely "optimization by relevance feedback" and "optimization by automatic term selection." These make the strategies practically workable as discussed in the last subsection in this section (D6).

5. Construction of Strategic Subqueries by Valency Control Strategy Application

Strategic subquery is the key structure for realizing the valency control strategy in the query. This subsection describes the method of how to arrange the terms and secondary indexing marks inside a strategic subquery based on a strategy. We have discussed and introduced a variety of secondary indexing terms for linguistic cues, summarized in section III.B.7. Using these secondary indexing markers and retrieval operators, designing an effective query construction based on a specific valency control strategy is relatively straightforward. The construction of a query formula corresponding to a verb-noun combination depends on the valency specification of the verb. (58) shows the general format and examples of queries of every valency specification variation. Each specification was classified as either lexical or syntactical, as labeled. These variations of expressions set the extent of valency controlled paraphrases. Note that the applicability of a strategy differs from verb to verb. For example, certain verb doublets may not have the short passive form, whereas might.

(58)

a. Transitive (-ϕ, -e) = Lexical
    #passage14(#1(noun @DO), verb_@ORG)
    e.g., #passage14(#1(空気 @DO) 汚す_@ORG)
    for "価格を上げる" 'X raises the price.'

b. Intransitive (-ϕ, -e) = Lexical
#passage14(#1(noun @SUBJ), verb_@ORG)
for "空気が汚れる" 'The air is contaminated.'

c. Short_Causative (-as) = Lexical or Syntactical
#passage14(#1(noun @DO) #2(verb_@NF @CAUS-s))
for "空気を汚す" 'X contaminates the air.'

d. Short_Passive (-ar) = Lexical
#passage14(#1(noun @SUBJ) verb_@ORG)
for "価格が上がる" 'The price rises.'

e. Short_Potential (-e) = Syntactical
#passage14(#1(noun @SUBJ) #2(verb_@NF @POTN))
for "価格が上げれる/上がれる" 'X can raise the price,' or 'the price can rise.'
#passage14(#1(noun @DO) #2(verb_@NF @POTN))
for "価格を上げれる" 'X can raise the price.'

f. Causative (-[s]ase) = Syntactical
#passage14(#1(noun @DO) #2(verb_@NF @CAUS-l))
for "価格を上げさせる/上がらせる" 'Y makes X raise the price,' or
'X makes the price rise.'

g. Passive (or Long Potential) (-[r]are) = Syntactical
#passage14(#1(noun @SUBJ) #2(verb_@NF @PASS-l))
for "価格が上げられる/上がられる" 'The price is raised,' 'X can raise the price,' or
'the price can rise.'
#passage14(#1(noun @DO) #2(verb_@NF @PASS-l))
for "価格を上げられる" 'The price is raised,' (indirect passive), or 'X can raise the price.'

h. Compound_Potential (V[adverbial]+e, or V+kotogaDekiru) = Syntactical
#passage14(#1(noun @SUBJ) #2(verb_@NF @POTN))
for "価格が上げれ得る/上がれ得る"
also for "価格が上げることができる/上がることができる"
'X can raise the price' 'price can rise.'

#passage14(#1(noun @DO) #2(verb_@NF @POTN))
for "価格を上げることができる"
also for "価格を上げれ得る"
'X can raise the price.'

i. VN_Causative (-s+as) = Syntactical
#passage14(#1(noun @DO) #2(verbal_noun @PREDICATE_VN @CAUS-w))
for "空気を汚染させる 'X makes the air be contaminated.'

j. VN_Passive (-s+are) = Syntactical
#passage14(#1(noun @SUBJ) #2(verbal_noun @PREDICATE_VN @PASS-w))
for "空気が汚染される 'The air is contaminated.'

k. VN_Potential (-[surukotoga]deki) = Syntactical
#passage14(#1(noun @SUBJ) #2(verbal_noun @PREDICATE_VN @POTN))
for "空気が汚染（することが）できる"
'X can contaminate the air,' or 'the air can be contaminated.'
#passage14(#1(noun @DO) #2(verbal_noun @PREDICATE_VN @POTN))
for "空気を汚染（することが）できる" 'X can contaminate the air.'

l. VN_Genitive (-no) = Syntactical
#2(noun @GEN_ACC/_NOM verbal_noun)
for "空気の汚染" 'contamination of air'

In a basic query construction, there are three steps to be carried out: First, the subject marker (@SUBJ) or the object marker (@DO) is coupled with the argument noun by a proximity operator (#proxN). Second, this coupled unit is placed in a #passage operator. Third, the valency specification (e.g., @PASS-l for the passivization) is also inserted in the #passage after the coupled argument unit.

Note that Japanese grammar allows the object of a potential transitive verb to take not only typically accusative -o (coded by @DO), but also typically nominative -ga (coded by @SUBJ). Note also that lexical formations, i.e., transitive with "-ϕ" or "-e" (52a), intransitive with "-ϕ" or "-e" (52b), and short passive with "-ar" (52d), are not accompanied by secondary markers like @TRANS, @INTRANS, etc., but modified by @ORG marker instead of @NF. This is because specifying the original verb form is enough to specify a lexical entity. However, in the case of the short causative (52c), there are cases of both lexical and syntactical formations. Hence, the supplementary attachment of the marker @CAUS-s onto the lexical item, transitivity or intransitivity being captured by @NF, is necessary as a syntactical operation. Among other syntactical cases (52f-h), unlike the previous two baselines with the @ORG marker, the normalized verb forms with @NF are adopted. This is because syntactic operation, in general, is highly productive and transparently applicable no matter what the transitivity of the verb is, if the operation does not carry some specific constraints, such as direct passivization for intransitive verbs. We intend here a simple automatic application of the method as desired in a practical system. Such a forceful, simple application of a strategic modification might also produce semantically unnatural sentence patterns. That is a cost of the adaptation of a general and simplistic solution, and greater sophistication of the system must be studied in our future work.

The last four formulae (52i-l) are designed for verbal nouns. They are also a straightforward mapping with operators of sentence (or phrase) patterns to the query term patterns. In the genitive case, the postpositional modifier phrase may express a relationship between the verbal noun and either the internal argument (as marked with @GEN_ACC) or the external argument (as marked with @GEN_NOM). We assign this marker simply based on the record in the segmentation dictionary - so if the verbal noun is classified as transitive (in the first description), the @GEN_ACC marker is assigned, and if it is intransitive, then the @GEN_NOM marker is assigned. Although this method is simple, and may not always give the correct class, at least a mismatch between the query and indexed text will never occur.

6. Optimization of Verb-Noun Combinations

The basic method of strategic subquery construction just described produces all possible verb-argument combinations. So, in some cases the number of the combinations may become very large. Worse, some combinations can have a harmful effect on retrieval performance because they may signify unintended contents. When whole possible verb-argument combinations were included in a single query, its performance fell significantly below the baseline. Thus, we need to somehow optimize the set of combinations of verbs and nouns in order to select positively effective verb-noun combinations. Two methods, 1) relevance feedback optimization, and 2) automatic optimization, are experimented with.

a. Optimization by Relevance Feedback

Relevance feedback is a method commonly used in information retrieval procedure to refine a query by analyzing the relevance of some selected (e.g., top N) retrieved documents. Similarly, the relevance feedback method for optimizing the selection of verb-noun combinations executes a pre-test of initial queries. However, our main purpose differs from that of the ordinary method - the goal of our specific relevance feedback procedure is to create a new sentence-level baseline for the prospective application of the valency control strategy in the next step. It is important for readers to understand this principle of our methodology. It is not to be confused with selecting the strategic subqueries themselves directly by measuring their performance.

Let us tentatively assume a very simple model of the valency strategy's effect on baseline performance. Let us say that a strategic subquery magnifies the gain or loss of the sentence-level baseline from the keyword-level baseline with a constant magnitude, if the sentence-level baseline subquery and the strategic subquery have identical sets of verb-noun pairs. Thus, a strategy is expected to polarize the effects of verb-noun combinations - a query of effective verb-noun combinations improves, and a query with an inadequate combination results in a worse performance. The insight of this speculation is as follows: Some verb-noun combinations may accidentally create a semantic bias toward contexts irrelevant to the true information need, so the strategic "effect" of wrong verb-noun combinations will eventually worsen performance, because it enhances the missignified semantics. However, instead of judging the "meaning" of a given verb-noun combination, we took a more formally definable approach consistent with the general attitude throughout this research. Thus, if we can collect a "good" set of verb-noun combinations (to make a new baseline) by some definable operational criteria, the performance gain by the strategy is expected to be boosted. Relevance feedback is one of the methods used to aim at such a goal. Thus, if a verb-noun combination is such a "good" one, its corresponding sentence-level baseline query will show some retrieval improvement from the keyword-level baseline, so that we can take its verb-noun combinations as the basis of succeeding strategic query experiments in the next step.

There is one technical dilemma in carrying out the above idea. On one hand, we may take every individual verb-noun pair in order to measure its effect. But in such a case the effect is elementary, and the impact on performance may be too small to measure. On the other hand, if we examine all the combinations of verb-noun pairs (i.e., in the power set of the set of pairs), the number of the combinations to be examined may become quite large (i.e., 2^(r*s), here r=<number of nouns>, s=<number of verbs>). Therefore, for practical reason, we need some method to set up some "reasonably" granular grouping of verb-noun combinations. Although there are numerous ways to do such a task, we chose the following method. If we have a verb-noun direct product, {N1, N2, ..., Nr}X{V1, V2, ...Vs}, we get its subset of r members, {{N1}X{V1, V2, ...Vs}, {N2}X{V1, V2, ...Vs}, ..., {Nr}X{V1, V2, ...Vs}}. Thus, each member contains a fixed noun element and whole set of corresponding verbs. Each {Ni}X{V1, V2, ...Vs} (i is 1,2, ..., r) is used to construct a constituent subquery to measure its performance. Thus, we create a set of pre-test queries for evaluation in which each query has a structure corresponding to Ni (0<i<=r) as shown in (59):

(59) #sum(<keyword-level baseline subquery>
#passage(N_i V_1) #passage(N_i V_2) ... #passage(N_i V_s))

We took this construction method because blending all possible verbs (which are key figures of the argument structure) in a single query will likely give an averaged effect of these verbs' strategic application.

In next step, we run these r queries and their retrieval performances are compared with the keyword-level baseline. Technically speaking, we set the selection criteria of "good" pre-test queries by judging their precision value at the low-end of recall. Thus, if their precision is higher than the corresponding value of the keyword-level baseline, the sentence-level baseline query is chosen, and otherwise, it is rejected.

Note that, consequently, this method usually produces more than one query from a single topic, and we measured all these varieties. Instead, if we choose only one group, e.g., the most improved one, we end up selecting only one argument noun and eliminating the rest. As a result, we will exceedingly limit the applicable domain in a given set of indexing terms. From the point of view of experimental design, this is quite undesirable. As we wrote, there are numerous approaches to set such groupings of verb-noun pairs, and ours represents only one of the ways to do it. Other operational principles must be examined in our future studies. Nevertheless, it should be emphasized that the relevance feedback procedure described here is designed and tested in experiments to measure the effect of strategic query control.

As the final step of the procedure, we have to generate the corresponding strategic queries based on the selected sentence-level subqueries. The operation of this step has already been described in the previous subsection (III.D.5).

At the end of this subsection, we show a sample case of selection of verb-noun combinations in (60). (60a) is an identical to (55b), which is a list of verb-noun direct products after the semantic augmentation of original keywords in topic j07 in our experiment. (60b) is a list of (two) accepted groups of verb-noun pairs, and (60c) shows (seven) rejected groups. Although the augmentation of initial keywords in the topic once expanded freely in the production process of keyword-level subquery (see D.3 in this chapter), the terms are again highly regulated by the relevance feedback method.

(60)

a. (=49b)
V×N = {{大気, 空気, 水質, 水, 河川, 湖水, 海洋}×{汚す/汚れる},
{ {atmosphere, air, water quality, water, river, lake, ocean}×{contaminate/be.contaminated},,
{ゴミ, 廃棄物}×{棄てる, 捨てる, 焼く/焼ける}}
{garbage, waste}×{dump, dump, burn(tr)/burn(intr)} }

b. (accepted verb-noun groups) =
1. { <空気, 汚す/汚れる> }
{<air, contaminate/be.contaminated> }
2. { <ゴミ, 捨てる>, <ゴミ, 棄てる>, <ゴミ, 焼く/焼ける> }
{ <garbage, dump>, <garbage, dump>, <garbage, burn(tr)/burn(intr)> }

c. (rejected verb-noun groups) =
1. { <大気, 汚す/汚れる> }
{<atmosphere, contaminate/be.contaminated> }
2. { <水質, 汚す/汚れる> }
{<water quality, contaminate/be.contaminated> }

3. { <水, 汚す/汚れる> }
{<water, contaminate/be.contaminated> }

4. { <河川, 汚す/汚れる> }
{<river, contaminate/be.contaminated> }

5. { <湖水, 汚す/汚れる> }
{<lake, contaminate/be.contaminated> }

6. { <海洋, 汚す/汚れる> }
{<ocean, contaminate/be.contaminated> }

7. { <廃棄 物, 捨てる>, <廃棄 物, 棄てる>, <廃棄 物, 焼く/焼ける> }
{ <waste, dump>, <waste, dump>, <waste, burn(tr)/burn(intr)> }

b. Automatic Optimization by Frequency Statistics

As we have just seen, the relevance feedback method for selecting verb-noun pairs to optimize their expected contribution had two practical problems. First and the foremost, it requires running the retrieval and analyzing the pre-test queries prior to the examination of targeted strategic queries. Second, the element-wise approach of picking up verb-noun pairs may not give reliable information for judging their qualification in the sentence-level baseline, which leads to a high gain in strategic queries in later application. Consequently, because of the large number of groupings of verb-noun pairs, we have to adopt a heuristic method of grouping, for which we so far do not have solid justification, for any specific method. Also, it may create more than one candidate query from a single topic. The automatic method that we are describing here does not have these shortcomings. If we could demonstrate improvement in a strategy's retrieval by using this automatic method, it would be meaningful for the practical application of our linguistic query strategy. Expecting to achieve a result comparable to (or at least not significantly lower than) the relevance feedback performance by this automatic method, we propose the following additional hypothesis to our set of hypotheses described previously (section III.A)[42] :

Hypothesis-RA [Relevance-Feedback.vs.Automatic]: The automatic method for the optimization of verb-noun combinations in query construction performs comparably to the relevance feedback method.

The procedure for automatic optimization of the selection of verb-noun pairs is quite simple: Choose high frequency pairs, and reject low frequency pairs from the query using some threshold. The reasoning behind this automatic method is as follows. A low frequency verb-noun pair in the strategic subquery affects the retrieval performance as a distraction factor. This means it works not only to degrade the recall, but also the precision. Let us assume that we have a set of documents retrieved by a query that contains all verb-noun pairs but only one pair in the strategic subquery, and this strategic pair appears rarely in the text collection. If we include this low frequency pair in the strategic subquery, two situations occur. One is the degradation of recall, because documents which were previously judged not relevant turn out to be judged as relevant. In other words, when verb-noun combinations specified in a query are found to be scarce in texts at retrieval time, performance degrades seriously. This is so even if the semantics of the verb-noun pair is perfectly relevant to the information need, because secondary markers such as @SUBJ, @TRANS, etc. in the query are found everywhere in the texts, and they get partial credit (of #passage operator) in the query evaluation. The other harmful effect is the degradation of precision. The ranking order of documents that were previously retrieved is

[42] Statistically speaking, we can show only the significance of the difference between two groups, but not their equivalence such as stated here. We can only satisfy this hypothesis by claiming it could not be rejected with a certain confidence level.

jeopardized because the solo distribution of the secondary markers is totally independent from the content words (i.e., primary indexing terms), and their frequencies (for example, in TF*IDF) become another major factor in determining the belief (or distance) value for relevance with the same mechanism as the recall case.

The following procedure was actually taken for this automatic optimization method. First, generate all possible (r*s [r is the number of argument nouns, s is the number of verbs]) combinations of verb-noun pairs, and measure the document frequency of each combination in the collection.[43] Second, all verb-noun combinations which have zero frequency are dropped, and the mean and standard deviation of the rest are calculated. Drop the verb-noun pairs which have frequencies outside from the range of the mean minus standard deviation up to the maximum. Finally, create a sentence-level baseline subquery and a corresponding strategic subquery based on the selected verb-noun pairs. Note that this automatic method does not utilize the relevance information, and it chooses a subset of verb-noun pairs element-wise. Therefore, no other alternative subset is chosen as a candidate for the prospective sentence-level baseline subquery. We show a sample case of selection of verb-noun combinations using this automatic optimization method in (61). (61a) is identical to (55b) which is the list of verb-noun direct products after the semantic augmentation of original keywords in the topic j07 in our experiment. (61b) is a list of (three) accepted verb-noun pairs, and (61c) shows (eight) rejected combinations.

[43] To measure this frequency value, we utilized the function of INQUERY which returns the number of retrieved documents. Thus, we run the query #14(<noun> <verb>) (here, 14 is the proximity window size of half of the average sentence size, equal to the passage window size for actual queries) for each verb-noun combination to collect the data. However, it should be noted that we did not use any relevance information. So, in an implemented practical system, such a measurement and an actual run of the strategic query can be done in a single step.

(61)

a. (=49b)
V×N = {{太気, 空気, 水質, 水, 河川, 湖水, 海洋}×{汚す/汚れる},
{ {atmosphere, air, water quality, river, lake, ocean}×{contaminate/be.contaminated},,
{ゴミ, 廃棄物}×{棄てる, 捨てる, 焼く/焼ける}}
{garbage, waste}×{dump, dump, burn(tr)/burn(intr)} }

b. (accepted verb-noun combinations) =
{ <空気, 汚す/汚れる>, <ゴミ, 捨てる>, <ゴミ, 焼く/焼ける>,
<廃棄 物, 捨てる> }
{<air, contaminate/be.contaminated>, <garbage, dump>, <garbage, burn(tr)/burn(intr)>,
<waste, dump> }

c. (rejected verb-noun combinations) =
{ <大気, 汚す/汚れる>, <水質, 汚す/汚れる>, <水, 汚す/汚れる>,
<河川, 汚す/汚れる>, <湖水, 汚す/汚れる>,<海洋, 汚す/汚れる>,
<廃棄 物, 棄てる>, <廃棄 物, 焼く/焼ける> }
{<atmosphere, contaminate/be.contaminated> , <water quality, contaminate/be.contaminated>,
<water, contaminate/be.contaminated>, <river, contaminate/be.contaminated>,
<lake, contaminate/be.contaminated>, <ocean, contaminate/be.contaminated>,
<waste, dump>, <waste, burn(tr)/burn(intr)> }

7. Summary of Query Formulation

Various methodological and technical issues about the query formulation have been discussed in this section. A method of query formulation is the locus where a theoretical model of valency control strategy of grammatical paraphrasing is implemented as actual query formulae in a specific retrieval system, in our case INQUERY. In other words, a strategically formulated query is a model-based realization under some given practical constraints or limitations, such as available retrieval operators or indexing representations.

The basic framework of the query was described as a three-level structure consisting of the keyword level, sentence level, and strategic level. A keyword level subquery corresponds to the most basic form of representation, which consists of all (primary) keywords with no structural arrangement. With this level of subquery formulation, the query is able to capture all possible "retrieved" documents with full recall. The sentence level subquery is designated to eliminate the bias of the sentence effect from the result of the targeted query strategy. These two levels of subqueries correspond to the two levels of baselines. The unbiased effect of grammatical paraphrasing is measured as a divergence of the performance from the sentence level baseline performance. In our methodology, these three subqueries are simply listed inside a #sum operator (which is, the most "neutral" among INQUERY operators) in order to make the query as straightforward as possible. Each strategic term pattern in a subquery is implemented by a combination of pattern specifiable operators, such as a proximity operator #proxN and a passage operator #passageN.

The last research issue we posed was a method to optimize the selection of verb-noun patterns from all their possible combinations. Once we select a suitable subset, we can construct a sentence-level baseline query and valency controlled strategic queries. We proposed two methods. One is the relevance feedback method, and the other is the automatic method. The relevance feedback method utilizes the relevance information of specific verb-noun combinations to choose a beneficial subset. The automatic method employs the cooccurrence information of a verb-noun pair. If it cooccurs more frequently than the threshold, the pair is selected. Thus, the automatic method does not depend on the relevance information, and is favorable if it performs as well as the relevance feedback method.

In the next chapter, we will present a series of experimental results which show the performance improvement achieved by the queries described in this section.

CHAPTER IV

EXPERIMENTS AND ANALYSIS

In this chapter, we will set forth various experimental results and analyses based on the hypotheses described in the previous chapter. As a logical sequence, first we will show the results of the baseline experiments, which include the contrasting of keyword-level and sentence-level baselines. Thus, these results will indicate the sentence effect. Our main experimental results and analyses follows. Both experimental results and analyses are arranged based on the hypotheses in our experimental design (see III.A).

As we have constantly emphasized, our main focus rests on a pair of dichotomies, i.e., the valency dichotomy (bivalent vs. monovalent) and the linguistic module dichotomy (lexicon vs. syntax). Hypotheses based on each dichotomy will be tested separately. Other results, such as those concerning the potential strategy and the verbal noun strategy, will be examined, too. The performance contrast between the relevance feedback and automatic methods is tested in order to measure the practicality of our method of strategic queries. A general assertion about the effectiveness of the valency control approach will be a part of the ultimate aim of this dissertation, in order to suggest future approaches to linguistic information retrieval.

A. Baselines and Sentence Effects

Figure 5 and Figure 6 (the data table is shown in Table 13 at the end of this section) show the keyword-level (labeled as "Base (KW)") and sentence-level (labeled as "Base (Sentence)" baselines for verbs given by the relevance feedback method and the automatic method, respectively. We also included the results of "normalized" sentence-level baselines (labeled as "Base (NF)") in these graphs. Here, the normalization means the stemming operation applied to transitivity alternation suffixes to demonstrate the effect of this specific morphology. (These suffixes are derivational, and as described before, all inflectional suffixes are unconditionally stripped from verbs in the process of word identification (see III.B.2).

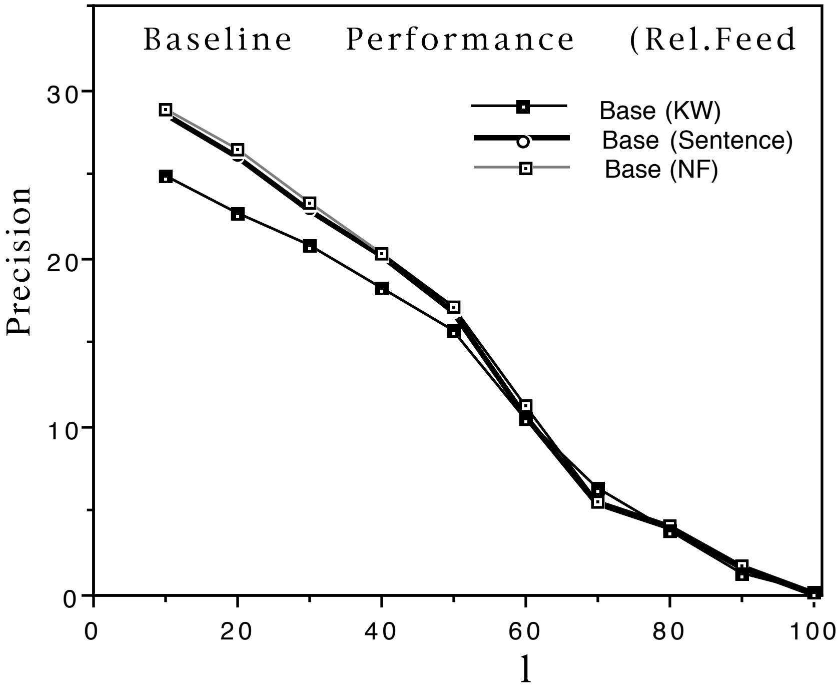


Figure 5. Keyword/sentence-level baselines (verb, relevance feedback).

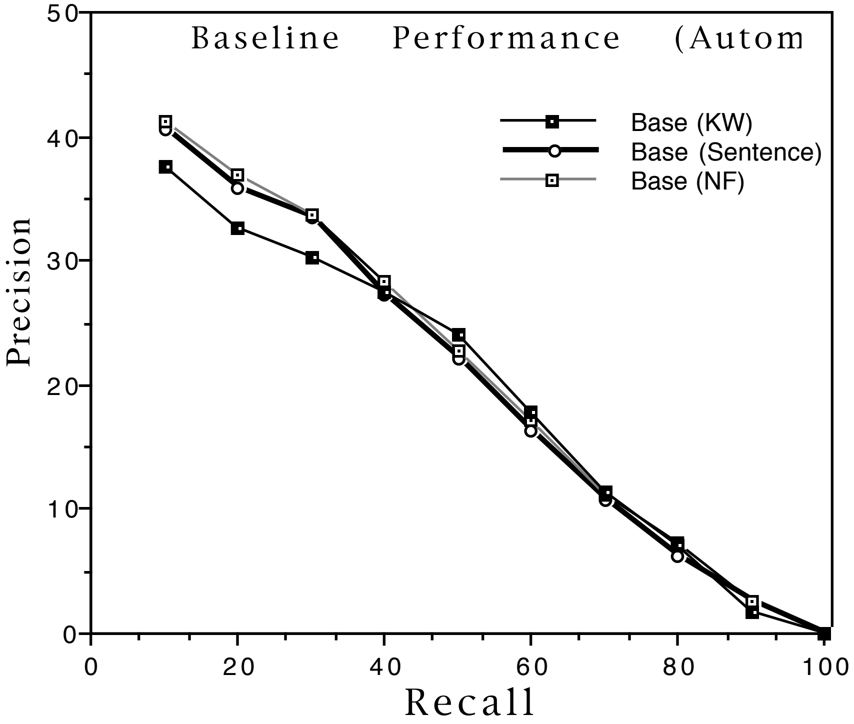


Figure 6. Keyword/sentence-level baselines (verb, automatic).

Both the relevance feedback and automatic methods show very similar behavior, even though their baselines differ from each other. In both cases, the precision values at the low recall level clearly improved (15.3% up (24.8% ->28.6%) in relevance feedback, 9.8% up (36.9%->40.5%) in automatic, at the low-end of recall). The effect of transitivity normalization had little impact on the performance in either case. Theoretically speaking, this normalization effect is not a part of the effect of the valency control strategy (for a verb-argument relationship), and should be separated from such strategic effects. However, because of this faint effect, we did not take this transitivity normalization into consideration when we analyze the effects of valency control strategies in our main experiments. (These experiments used non-normalized verb forms.)

Figure 7 and 8 are the baseline results for verbal nouns. (The data table is shown in Table 13 at the end of this section) These results do not contain data for transitivity normalization because verbal nouns cannot alternate their transitivity morphologically. In the relevance feedback case (Figure 7) the curves are sluggish, and improvement of the sentence effect was small (8.5% up (21.3%->23.1%)). In contrast, the automatic method (Figure 8) showed smooth and consistent improvement (11.9% up (32.7%->36.6%)).

Thus, we got consistent results of precision improvement at the low recall end in both verbs and verbal nouns. This is a clear indication of the sentence effect as a precision enhancement device. The results of the sluggishness of relevance feedback and the smoothness of the automatic technique suggest greater reliability of the automatic than of the relevance feedback method.

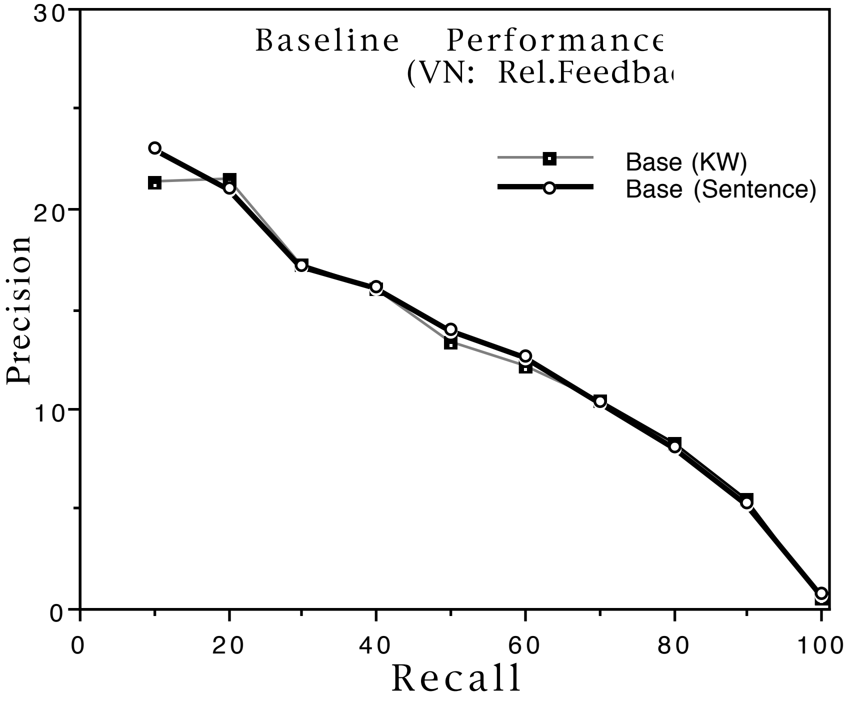


Figure 7. Keyword/sentence-level baselines
(VN, relevance feedback).

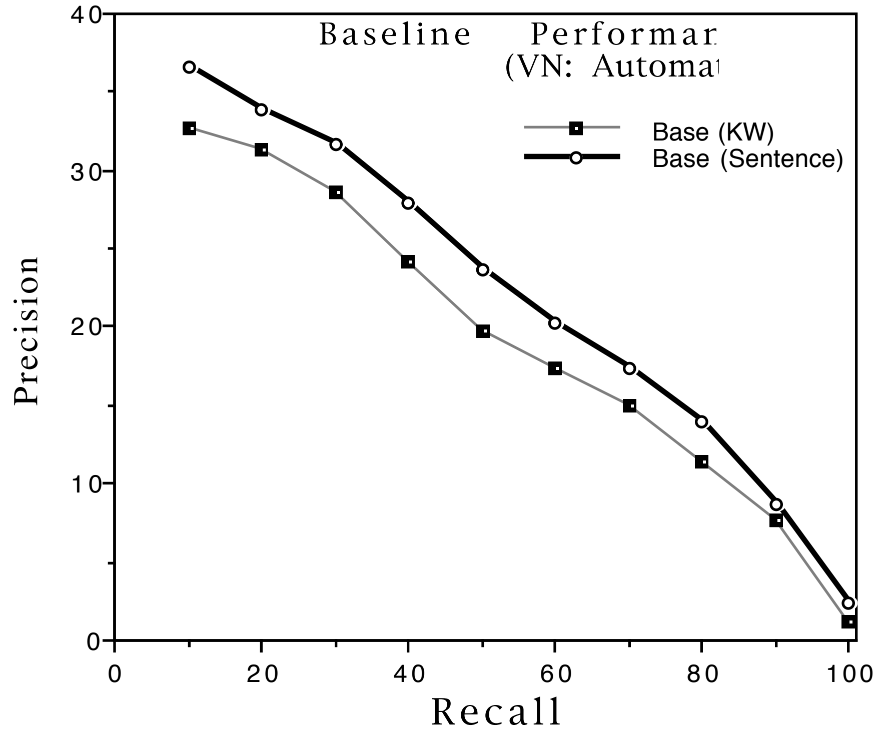


Figure 8. Keyword/sentence-level Baselines (VN, automatic).

Table 13. Results of baseline performance.

[Relevance Feedback]:

Recall	V: KW	V: Sentence(%Inc)	V: NF(%Inc)	VN: KW	VN: Sentence(%Inc)
10	24.8	28.6(+15.3)	28.8 (+16.1)	21.3	23.1 (+8.5)
30	20.8	23.0(+10.6)	23.3 (+12.0)	17.2	17.2 (0.0)
50	15.6	16.9 (+8.3)	17.1 (+9.6)	13.3	14.0 (+5.3)
80	3.8	4.1 (+7.9)	4.1 (+7.9)	8.3	8.2 (-1.2)
100	0.1	0.1 (0.0)	0.1 (0.0)	0.5	0.8(+60.0)

[Automatic]:

Recall	V: KW	V: Sentence(%Inc)	V: NF(%Inc)	VN: KW	VN: Sentence(%Inc)
10	36.9	40.5 (+9.8)	41.3 (+11.9)	32.7	36.6 (+11.9)
30	30.5	33.4 (+9.5)	33.7 (+10.5)	28.6	31.6 (+10.5)
50	20.8	22.1 (+6.2)	22.7 (+9.1)	19.8	23.7 (+19.7)
80	7.0	6.2 (-11.4)	7.2 (+3.0)	11.4	13.9 (+21.9)
100	0.1	0.1 (0.0)	0.1 (0.0)	1.2	2.4 (+50.0)

B. Experimental Results

We set forth the experimental results of various valency control strategies as follows. These basically correspond to the facets of the experimental design that was schematized in Table 2 in the first section of this chapter (III.A). Note that some experiments with verbs have fewer than a full set of queries (for verbs, 9 queries (5 topics) for relevance feedback, and 6 queries (6 topics) for automatic; for verbal nouns, 24 queries (6 topics) for relevance feedback , and 6 queries (6 topics) for automatic). This is because verbs in these queries do not have the transitive alternative forms that are applicable to the corresponding query strategies. In such a case, they consequently have different baselines, and are not precisely comparable to other experiments with different baselines. They are noted in each description of experimental result, as necessary.

These queries (from 6 topics) include 32 unique verbs (17 transitive (but not short-causative) verbs, 5 intransitive (but not short-passive) verbs, 3 short-causative verbs, 7 short-passive verbs), and 17 verbal nouns. For intransitives including short-passives, the only possible argument nouns are at the subject position (i.e., @SUBJ marker), though transitives including short-causatives have two alternative designations, that is at the subject (@SUBJ) and object (@DO) positions. Of 17 transitive (but not short-causative), 8 are non-doublet verbs, and of 5 intransitive (but not short-passive), 1 was a non-doublet verb. There are no cases of non-doublets in short-causatives and short-passives. Thus, 6 verbs out of a total of 32 are not doublet verbs, and lexical (i.e., transitive and intransitive) valence control strategies are obviously not applicable to these non-doublet verbs. (Also see Table 2 in section II.D for the statistics.)

We are expecting the general effects of valency control strategies to show themselves as a precision enhancement device. In other words, the precision-recall curve will indicate some gain at the low recall level, but less gain or possibly slight degradation at the mid to high recall level. The following report will present and focus on the precision gain at the low end of recall as numbers, as well as the precision-recall curve as a graph.

1. Transitive vs. Intransitive

The retrieval performance with transitive and intransitive strategies is shown in Table 14 for both the relevance feedback and automatic experimental sets. These data are presented as graphs in Figures 9 and 10, respectively. This experiment includes 8 queries (5 topics) in relevance feedback, 6 queries (6 topics) in automatic.

In the case of transitives, the curves clearly indicate some gain in the low recall region (<40%), and degradation at the middle to high recall region (>40%), in both relevance feedback and automatic experiments. The relevance feedback performance gain at the low recall end was +12.2% (28.6% ->32.1%), and the automatic method's improvement was +9.6% (40.5% ->44.4%).

In the intransitive experiments, such consistent characteristics are not clearly observed. Although the curve of the relevance feedback experiment showed a similar tendency to the transitive, the result of the automatic method showed no improvement of retrieval effectiveness at any recall level.

Table 14. Retrieval performance of transitive and intransitive.

[Relevance Feedback]:

Recall	Intrans (%Inc)	Base(intr)	Trans (%Inc)	Base(trans)
10	31.5(+12.5)	28.0	32.1 (+12.2)	28.6
30	24.9 (0.0)	24.9	24.0 (+4.3)	23.0
50	16.9 (-7.7)	18.3	13.9 (-17.8)	16.9
80	3.5 (-10.3)	3.9	3.2 (-22.2)	4.1
100	0.1 (0.0)	0.1	0.1 (0.0)	0.1

[Automatic]:

Recall	Intrans (%Inc)	Trans (%Inc)	Baseline
10	38.3 (-5.4)	44.4 (+9.6)	40.5
30	31.0 (-7.2)	34.6 (+3.6)	33.4
50	22.7 (+2.7)	19.5 (-11.8)	22.1
80	8.3(+33.9)	5.8 (-6.5)	6.2
100	0.1 (0.0)	0.1 (0.0)	0.1

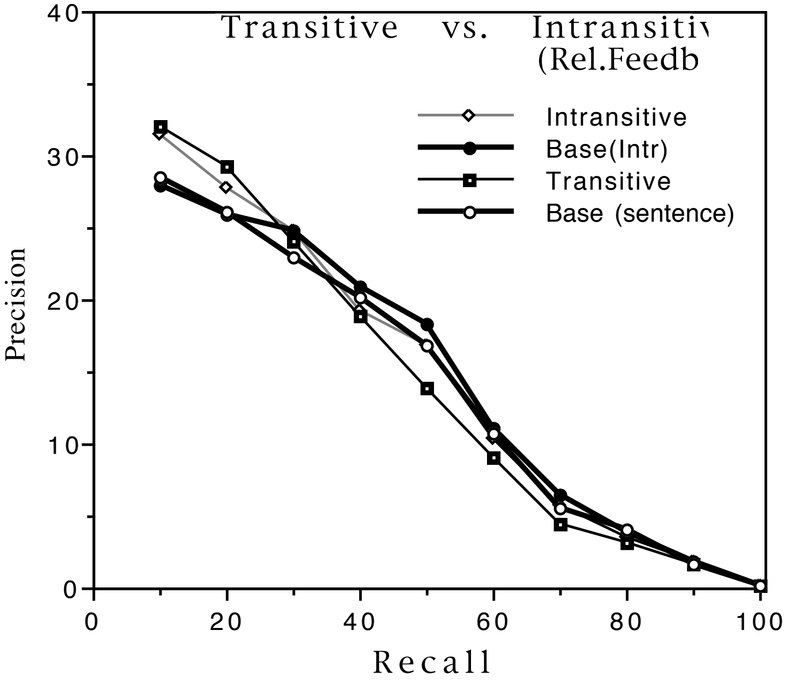


Figure 9. Retrieval performance by transitive strategy (relevance feedback).

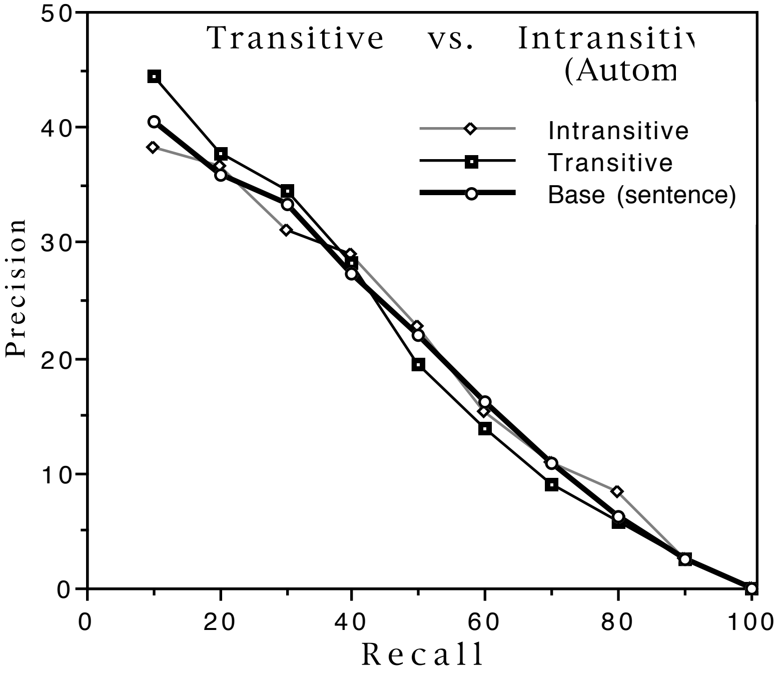


Figure 10. Retrieval performance by transitive strategy (automatic).

2. Causative vs. Passive vs. Potential

The retrieval performance with causative, passive, and potential strategies is shown in Table 15 for both the relevance feedback and automatic experimental sets. These data are depicted in graphs in Figure 11 and Figure 12, respectively. This experiment includes 8 queries (5 topics) in relevance feedback, 6 queries (6 topics) in automatic. We notice first that, as with the results of transitive and intransitive in lexical categories, there is a general tendency showing precision enhancement across all three constructions. The order of performance at the low recall end among the three strategies was causative, potential, then passive in both the relevance feedback and automatic cases. The difference was significant in the relevance feedback case, but subtle in the automatic case. Furthermore, in the automatic causative result, the performance at middle recall level was able to maintain the baseline performance. However, despite the moderate gain at low recall, the automatic potential case also showed some performance degradation at middle recall region. Thus, we may characterize the potential performance as somehow intermediate between causative and passive - improvement similar to causative, and degradation similar to passive. On the other hand, similar to the intransitive results in the previous experiments, the performance of passive, which is also a monovalent construction, showed poor retrieval performance. In the relevance feedback case, the performance is worse than the baseline at any recall level. In the automatic case, the performance is the worst of all.

Table 15. Retrieval performance of causative, passive and potential.

[Relevance Feedback]:

Recall	Causative (%Inc)	Passive (%Inc)	Potential (%Inc)	Base(sentence)
10	35.7 (+24.8)	27.4 (-4.2)	32.2 (+12.6)	28.6
30	23.5 (+2.2)	21.3 (-7.4)	19.8 (-13.9)	23.0
50	16.3 (-3.6)	12.6 (-25.4)	12.3 (-27.2)	16.9
80	3.7 (-9.8)	3.4 (-17.1)	3.3 (-19.5)	4.1
100	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.1

[Automatic]:

Recall	Causative (%Inc)	Passive (%Inc)	Potential (%Inc)	Base(sentence)
10	44.4 (+9.6)	42.2 (+4.2)	43.1 (+6.4)	40.5
30	34.4 (+3.0)	31.7 (-5.1)	34.8 (+4.2)	33.4
50	22.6(+2.3)	21.5 (-2.7)	20.3 (-8.1)	22.1
80	6.4 (+3.2)	8.6 (+38.7)	7.2(+16.1)	6.2
100	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.1

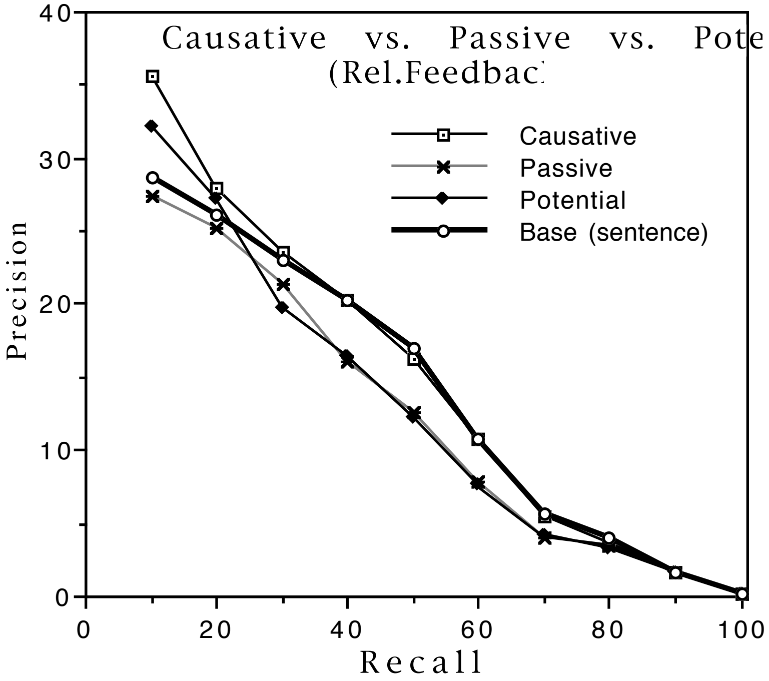


Figure 11. Retrieval performance by valency strategies (relevance feedback).

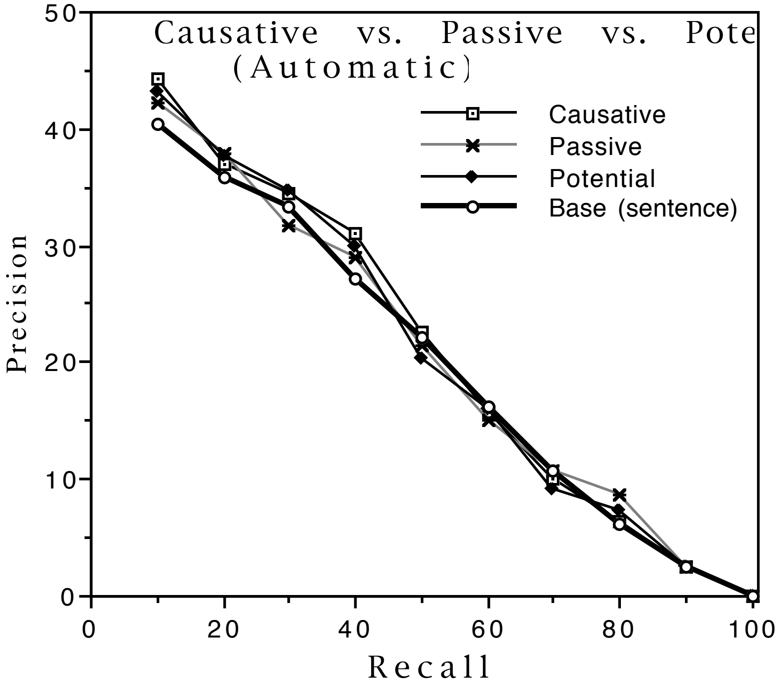


Figure 12. Retrieval performance by valency strategies (automatic).

3. Causative vs. Transitive and Passive vs. Intransitive

This data shows contrasts with that from a set of two previous experimental results (a. and b.). Thus, previous reports compared experiments across various valences within the same linguistic module. Information here indicates the reverse - distinct linguistic modules with an identical valence value. Since we have already described the numerical data in previous tables, we exhibit only the P-R curves. Figures 13 and 14 present the bivalent contrast i.e., between transitive and causative, and Figure 15 and 16 represent the monovalent construct, i.e., between intransitive and passive query constructions, in the relevance feedback mode and the automatic mode, respectively for both pairs.

In bivalent contrasts, for any condition, no matter whether lexical (i.e., transitive) or syntactical (i.e., causative), or whether by relevance feedback or automatic, some performance gain at low recall level was observed. Furthermore, transitive cases showed some performance degradation - a typical situation of precision enhancement by paying some price in recall. But interestingly, experimental results showed that causative cases did not indicate a gain drop from the baseline even at the middle to high recall level in either the relevance feedback and or the automatic case. In contrast, monovalent results were generally poor (especially relevance feedback passive), exhibiting an insignificant departure from the baseline (especially in automatic results), and inconsistency (such as the performance orders between relevance feedback and automatic).

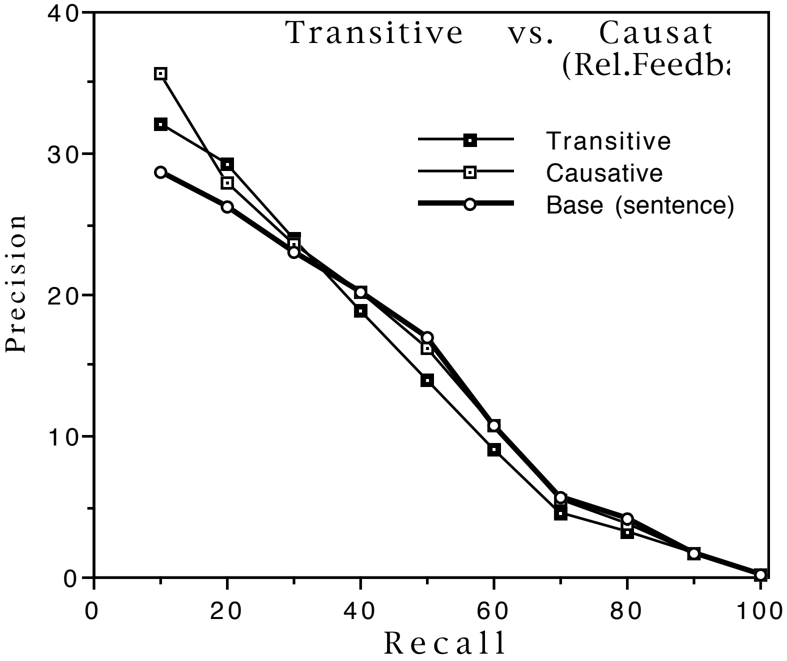


Figure 13. Retrieval performance by bivalent strategies (relevance feedback).

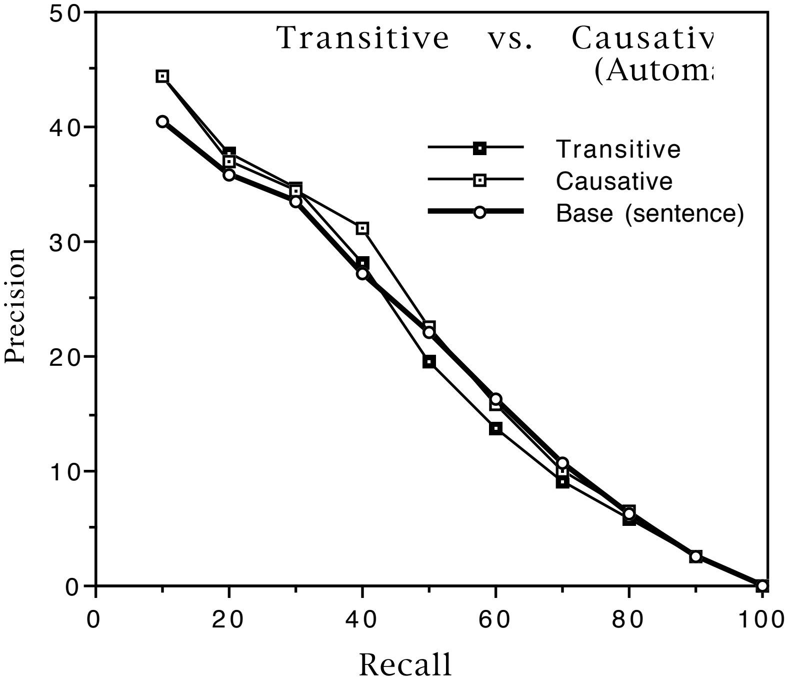


Figure 14. Retrieval performance by bivalent strategies (automatic).

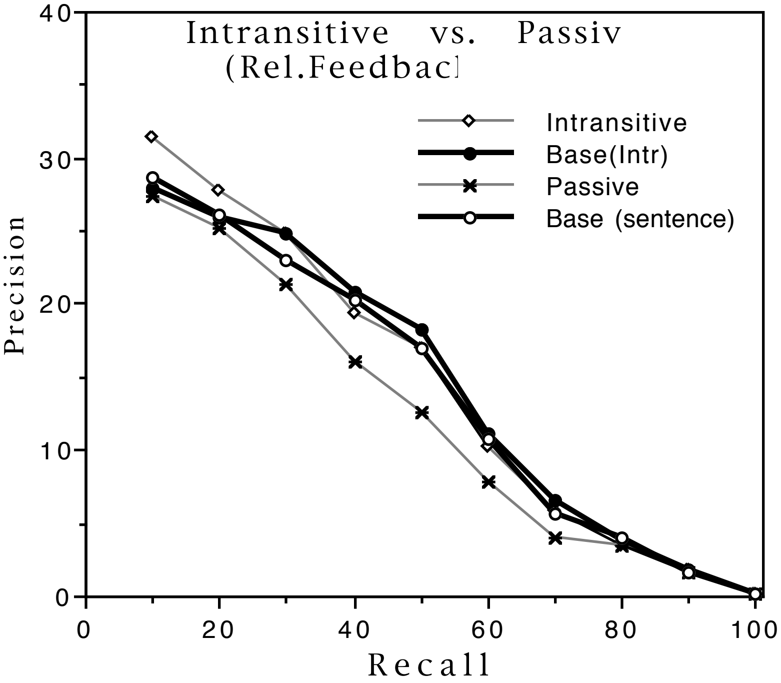


Figure 15. Retrieval performance by monovalent strategies (relevance feedback).

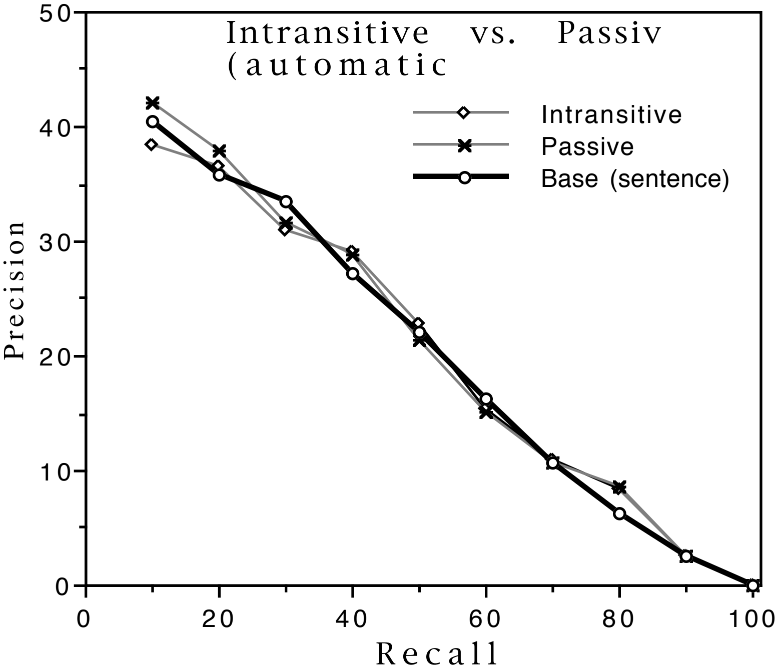


Figure 16. Retrieval performance by monovalent strategies (automatic).

4. Contrast on Combined Measures

To draw a more general picture of performance differences of strategic approaches in a single dichotomous view, we constructed a group of queries which share a particular dichotomous feature by combining corresponding strategies. That is, a "lexical" strategy consists of transitive and intransitive; "syntactical" is causative plus passive; "bivalent" is transitive plus causative; and "monovalent" is intransitive plus passive. The performance results are described in Table 16, and drawn in the graphs in Figure 17 and Figure 18, for relevance feedback and automatic modes, respectively. We recognize several characteristics from these graphs.

Some have the same interpretations as we noticed before. Thus, in most cases, we see an improvement in precision at the low recall region, except in the case of monovalent strategies, which show degradation throughout the levels of recall. In the relevance feedback results, there are performance drops at the mid-recall region, while they experienced precision gain at low recall. In the automatic data also, we can see a similar tendency, though the results are less clear than those for relevance feedback.

A noticeable new point here is that performance differs both between lexical and syntactical, and between bivalent and monovalent. Thus, we can see a significant performance difference between bivalent and monovalent results, as is logically understandable from the observation in the previous paragraph about the ineffectiveness of the monovalent strategies. However, the curves of the lexical and syntactical strategies are fairly close to each other in both the relevant feedback and automatic cases.

Table 16. Retrieval performance of combined strategies.

[Relevance Feedback]:

Recall	Lexical (%Inc)	Syntax (%Inc)	Bivalent (%Inc)	Monovalent	Base(sentence)
10	32.4 (+13.3)	31.6 (+10.5)	35.7 (+24.8)	29.0 (+1.4)	28.6
30	21.9 (-4.8)	21.9 (-4.8)	21.0 (-8.7)	20.1 (-12.6)	23.0
50	13.5 (-20.1)	13.9 (-17.8)	13.6 (-19.5)	12.6 (-25.4)	16.9
80	3.2 (-22.0)	3.3 (-19.5)	3.1 (-24.4)	3.4 (-17.1)	4.1
100	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.1

[Automatic]:

Recall	Lexical (%Inc)	Syntax (%Inc)	Bivalent (%Inc)	Monovalent	Base(sentence)
10	46.7 (+15.3)	44.1 (+8.9)	45.6 (+12.6)	36.9 (-8.9)	40.5
30	35.7 (+6.9)	33.8 (+1.2)	35.5 (+6.3)	30.5 (-8.7)	33.4
50	20.4 (-7.7)	21.1 (-4.5)	20.0 (-9.5)	20.8 (-5.9)	22.1
80	6.8 (+9.7)	8.3 (+33.9)	5.5(-11.3)	7.0 (+12.9)	6.2
100	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.1

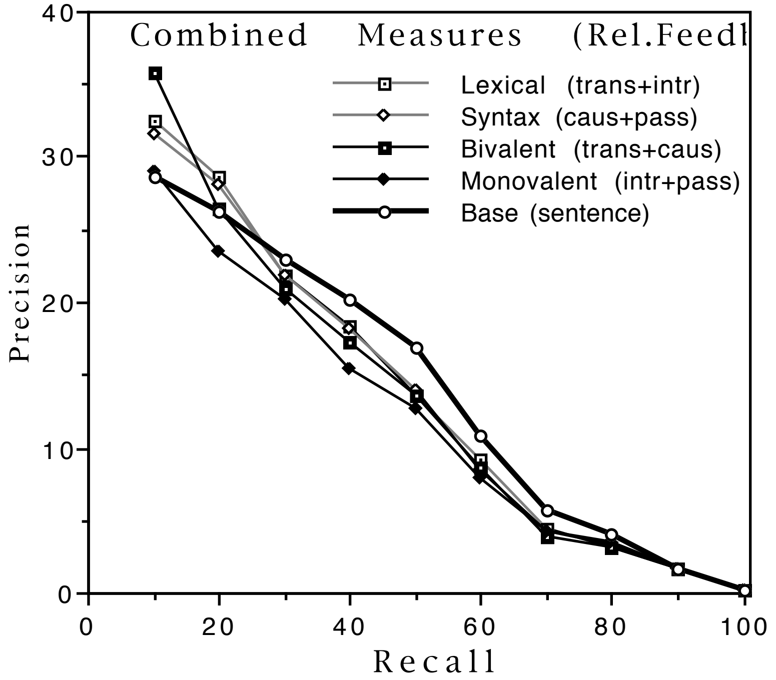


Figure 17. Retrieval performance by combined strategies (relevance feedback).

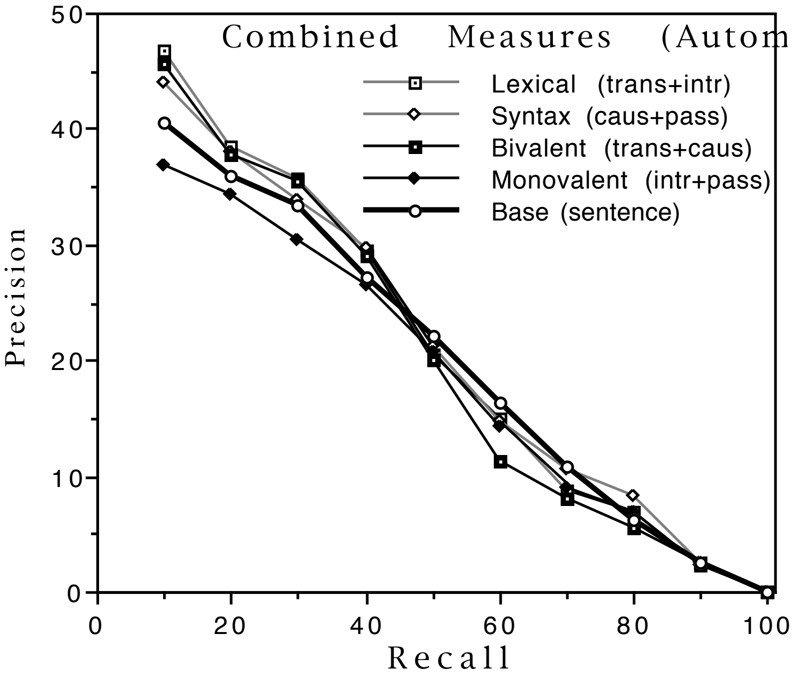


Figure 18. Retrieval performance by combined strategies (automatic).

5. Verbal Nouns

Table 17 shows the experimental results of retrieval performance by verbal noun query strategies. Figure 19 and Figure 20 depict these results graphically for relevance feedback and automatic data, respectively. Since Japanese verbal nouns are quite different from verbs both in grammatical function and in lexical semantics, their performances are expected to differ distinctively from each other.

As seen in the verb experiments, we can observe precision enhancement, i.e., performance improvement, at the low recall level for most verbal noun constructions, except in the case of genitive queries. The retrieval performance by genitive strategies was discouraging - queries with genitive formatives did not show a sign of improvement at any recall level in either relevance feedback or automatic experiments. This has some resonance with past research endeavors on the effective use of noun phrases (see Fagan (1987), among others), which are mostly either not encouraging or inconclusive at best. Genitive constructions have, in general, less specificity in their contextual semantics than full-scale clause constructions, because the simple nominal sequence linked by simple connectors hides grammatical and contextual relationships by "collapsing" them into such a simplified form with a covertly specified context.

Another distinctive characteristic of the verbal noun results is that the relevance feedback data made little difference, either for better or worse, from the baseline, while the automatic method showed wider range of deviation from the baseline. This situation contrasts with most data from verb experiments, which exhibits that relevance feedback data has more deviation than that of automatic methods. It is not clear whether this is simply a coincidental result arising from certain queries, or not. We need further experiments in the future to confirm this proposition.

Table 17. Retrieval performance of verbal nouns.

[Relevance Feedback]:

Recall	Caus:VN (%Inc)	Pass:VN (%Inc)	Potn (%Inc)	Gen (%Inc)	Comb (%Inc)
10	23.1 (0.0)	24.9 (+5.1)	25.4 (+10.0)	22.6 (-2.2)	25.3 (+9.5)
30	18.2 (+5.8)	18.4 (+4.5)	18.5 (+7.6)	16.8 (-2.3)	16.6 (-3.5)
50	13.5 (-3.6)	14.4 (+0.7)	13.2 (-5.7)	13.8 (-1.4)	12.8 (-8.6)
80	8.3 (+1.2)	8.5 (-1.2)	8.0 (-2.4)	8.2 (0.0)	7.1 (-13.4)
100	0.8 (0.0)	0.9 (0.0)	0.8 (0.0)	0.8 (0.0)	0.8 (0.0)

Recall	Base:VN	Base(Pass):VN
10	23.1	23.7
30	17.2	17.6
50	14.0	14.3
80	8.2	8.6
100	0.8	0.9

Table 17. Retrieval performance of Verbal Nouns (cont.)

[Automatic]:

Recall	Caus:VN (%Inc)	Pass:VN (%Inc)	Potn (%Inc)	Gen (%Inc)	Comb (%Inc)
10	38.6 (+5.5)	42.2 (+4.7)	47.0 (+28.4)	34.2 (-6.6)	45.3 (+23.8)
30	33.8 (+7.0)	37.3 (+4.2)	33.4 (+5.7)	29.5 (-6.6)	33.4 (+5.7)
50	24.8 (+4.6)	28.0 (+7.3)	24.2 (+2.1)	23.1 (-2.5)	26.0 (+9.7)
80	15.2 (+9.4)	17.1 (+5.6)	15.4 (+10.8)	13.9 (0.0)	14.1 (+1.4)
100	2.4 (0.0)	2.6 (0.0)	2.3 (-4.2)	2.4 (0.0)	2.3 (-4.2)

Recall	Base:VN	Base(Pass):VN
10	36.6	40.3
30	31.6	35.8
50	23.7	26.1
80	13.9	16.2
100	2.4	2.6

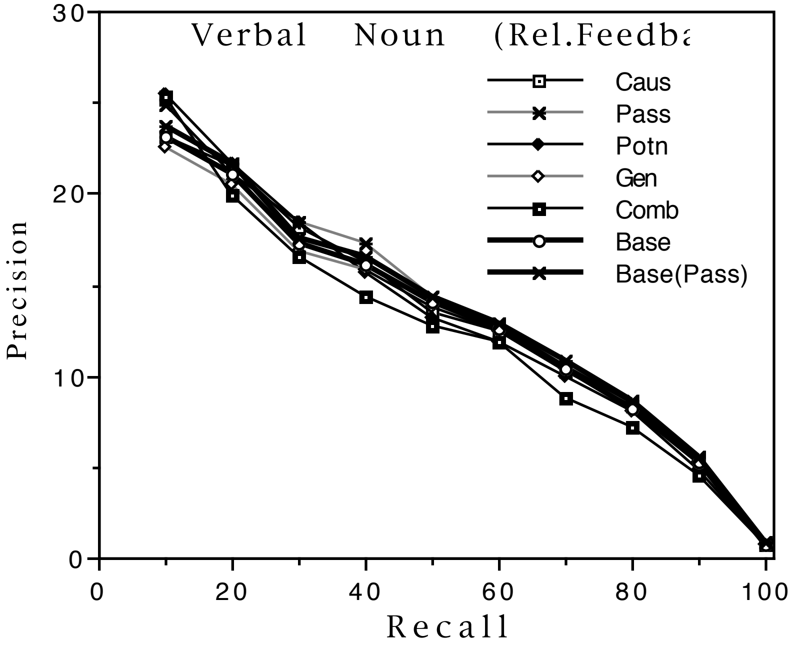


Figure 19. Retrieval performance by VN strategies (relevance feedback).

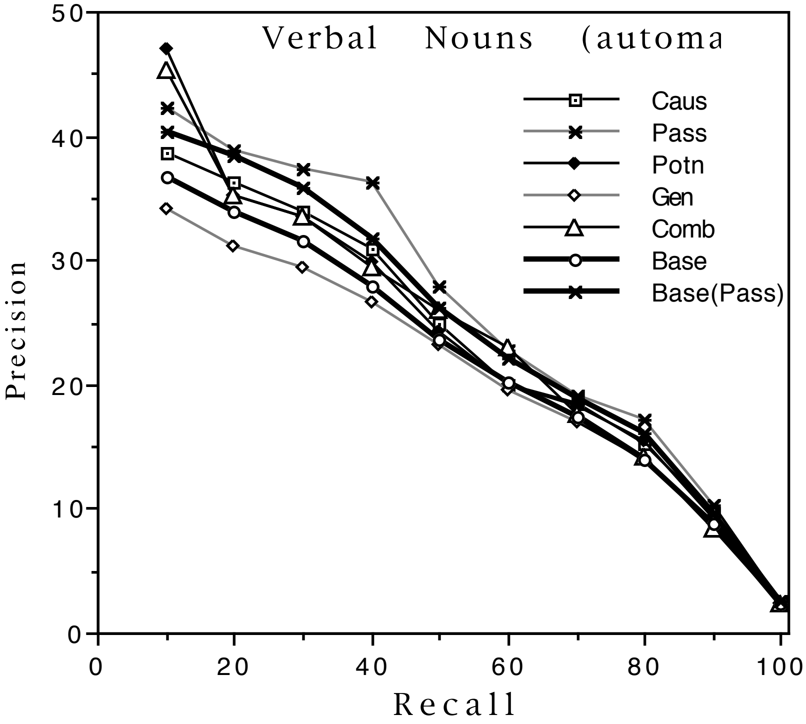


Figure 20. Retrieval performance by VN strategies (automatic).

C. Performance Analyses

1. General Review

Table 18 is a summary of the performance data of our retrieval experiments corresponding to our experimental design, which was previously depicted in Table 2 (section III.A). Since we focused on the improvement of the precision, the data are obtained from the low recall region, which was calculated as an average of two precision data at the 10% and 20% recall levels (i.e., 67 and 135 partial relevant documents in total 665 relevant documents, respectively). Figure 21 and Figure 22 are the visual depictions of Table 18 as histograms for verbs and verbal nouns, respectively.

Table 18. Experimental result summary (relevance feedback/automatic).

<baseline>	Syntax	Lexical	Combined	VN
Passive/Intr.	-4.0/+4.8	+9.7/-1.8	-4.7/-6.6	+1.2/+2.8
Causative/Trans.	+15.9/+6.4	+12.0/+7.3	+13.0/+9.1	+1.0/+6.4
Combined	+8.9/+7.5	+11.4/+11.2	---	+1.9/+13.8
Potential	+8.2/+5.6	---	---	+6.2/+16.3
Genitive	---	---	---	-2.5/-7.3

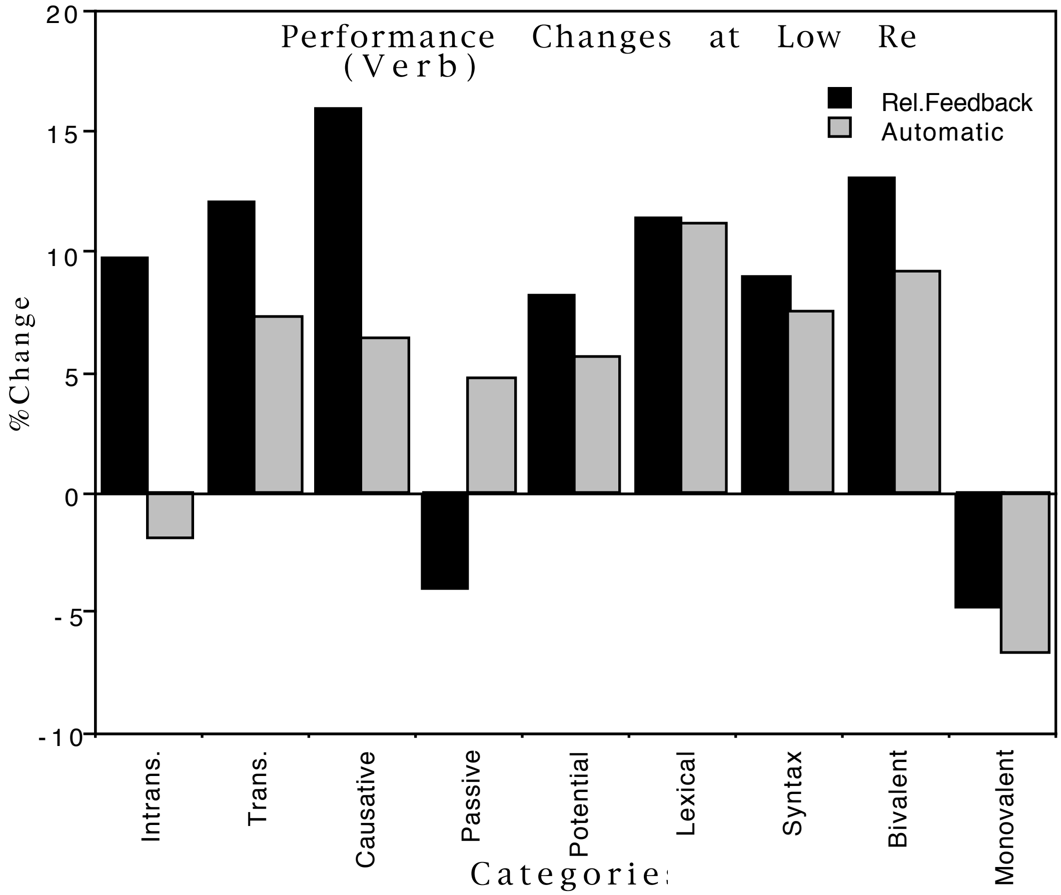


Figure 21. Summary of precision enhancements by verbal query strategies.

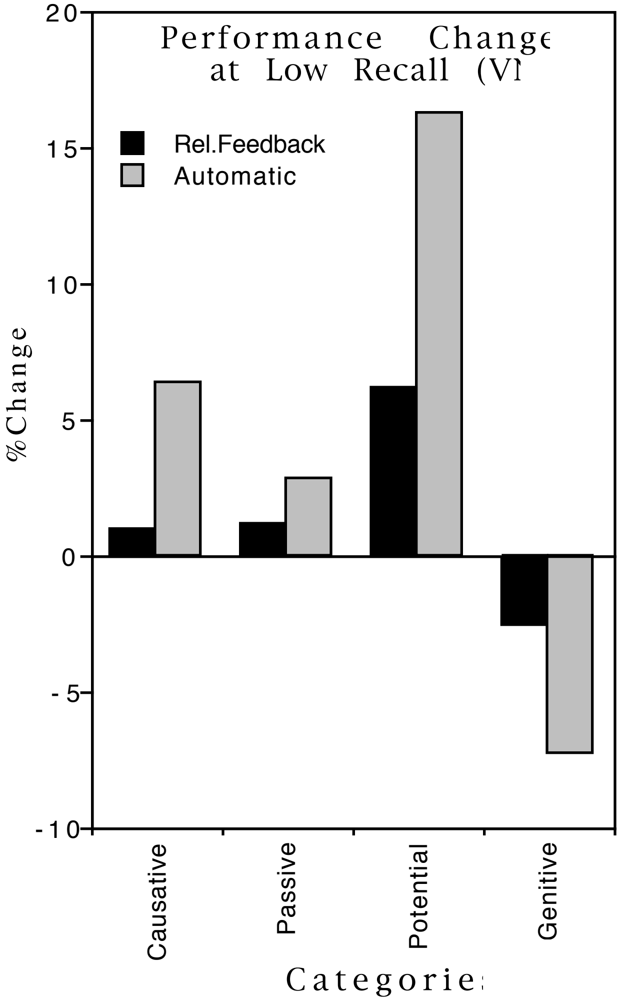


Figure 22. Summary of precision enhancements by VN query strategies.

Let us point out several distinctive characteristics of these data. First and most importantly, most valency-based query strategies improved precision from around 5% to 15%, except in only a few monovalent cases. Thus, we can claim the general effectiveness of valency control for the query as stated in Hypothesis-G (hereafter, we follow the names of the hypotheses as given in III.A). This is a positive evidence of the effectiveness of the application of natural language techniques to information retrieval.[44] We believe that this claim should be seen as be one of the major contributions of this dissertation, because of the hand-waving results of many natural language techniques in past information retrieval studies.

At a more detailed level, three approaches, i.e., the lexical, syntactic, and bivalent methods (Hypothesis-LEX, Hypothesis-SYN, and Hypothesis-BI), showed positive gain, and the monovalent method (Hypothesis-MONO) degraded effectiveness. For verbal nouns, all techniques, except the genitive one, also produced performance gains (Hypothesis-VN). As already described in the discussion about the experimental results (B.6), the automatic method applied to verbal nouns exhibited greater gains than the relevance feedback method (though it also showed a larger loss in the genitive case).

Individual elemental techniques seem to display more inconsistency than the combined techniques mentioned above. However, there is a clear tendency such that a bivalent strategy, i.e., the causative and transitive methods, exhibited good (from 6% to 16%) performance improvement. The results of monovalent strategies, i.e., passive and intransitive data, swung positively and negatively from the baseline, depending on the technique (i.e., intransitive or passive), and also on the optimization method (i.e., relevance feedback or automatic). It is a sign of the instability of this method. The potential method also showed moderate (from 6% to 8%) improvement.

[44] In the following discussion, we did not strictly apply statistical tests, unless so noted. There are two reasons for this. First, the methodology we adopted has a disadvantage such that we are aiming to achieve precision improvement only at the low recall level (10-20%), but no improvement or possible degradation from the middle to high recall levels. Thus, the applicable region for sign test, which is usually performed in similar experiments, is very narrow, and the power of judgment should be too weak. The T-test is not a well-suited method because the normality of the distribution cannot be assumed (Salton, p.171, 1983). Second, we were able to use only a limited number of queries. Thus, we have to wait for a larger scale experimentation to make our points firmer - it should be in our future work.

Interestingly, this intermediate result between causative and passive performance accords with its valence characteristic of no change (0) between that of causative (+1) and passive (-1). However, the potential strategy of verbal nouns achieved the best in its group.

In following sections, we will contrast the experimental results in various ways.

2. Analysis of Valency Dichotomy (Monovalent vs. Bivalent)

This performance analysis contrasts monovalent and bivalent query strategies, and corresponds to the Hypothesis-BM [Bivalent vs. Monovalent] (and three separate hypotheses, Hypothesis-TI [Transitive vs. Intransitive] limited in lexicon, Hypothesis-CP [Causative vs. Passive] limited in syntax, and Hypothesis-VNCP [VN Causative vs. VN Passive] limited in (syntactical) verbal noun constructions). We have already noted that any bivalent method generally exhibited good performance. This salient characteristic becomes very distinctive when we see the dichotomy in the combined categories, i.e., monovalent (=intransitive+passive) and bivalent (=transitive +causative). The bivalent method clearly demonstrated superior retrieval performance to the monovalent method in cases corresponding to all four hypotheses mentioned above. This is a valuable finding, because this claim is, linguistically speaking, formally definable rather than being dependent on semantic or pragmatic characterization, either of which tends to make retrieval an empirical and ad hoc art. A simple and straightforward explanation is that bivalency has more power to allow for contextual information than monovalent construction. Following the same line of thought, causatives always add some new information (e.g., causer) to the original, while passivization does not. This answer certainly "pushes the right button" to illustrate a principal mechanism of the role of valency information in retrieval. However, it seems insufficient, and we probably need further elaboration. Let us discuss the reason in the following ways:

First, we may point out the following counter-argument to the above notion of "inequality of the context information": A monovalent sentence can, at least theoretically, carry the same information that the corresponding active bivalent sentence does. Active and passive sentences are usually interchangeable and are transformable as in (56a, 56b). If we allow ourselves to construct a sentence even more freely, even an intransitive sentence can signify a similar meaning, as in (56c). Since the designation of the direct object, as opposed to the subject, is most common in a transitive query (at least in our experiments), the passive or intransitive counterpart has structurally no disadvantage in this case.

(56) a. Sono Otokonoko-ga Mado-o War-ta. (r-t=>tt)
(The boy broke a window.)
b. Mado-ga Sono Otokonoko-niyotte War-are-ta.
(A window was broken by the boy.)
c. Mado-ga Sono Otokonoko-no-seide War-e-ta.
(A window broke through the boy('s action).)

To consider this argument, let us look at some statistical data. According to Svartvik (1966, pp.140-143), 80% of English ("agentive") passive[45] clauses are "agentless," and the rest, 20% are "agentful." Let us

[45] In his classification (p. 9), passive sentences are primarily classified into *agentive* and *non-agentive*. According to his data (p.155), the agentive group is a majority (2696 cases) over the non-agentive group (666 cases). Agentive passive sentences are further classified into *agentful* and *agentless*.

assume a similar distribution of Japanese passive verbs. On the other hand, in Table 12 (section III.C), we noted that 12.6% of Japanese sentences are in the passive voice (i.e., PASS-l or PASS-w), and the rest, 87.4% are active. Thus, the majority of total clauses likely have an agent as in the active voice, but in the population of passives, the majority are agentless. This contrast suggests that there is a linguistic mechanism which controls our distinctive use of either voice - when there is a "need" to express an agent explicitly, we choose active, otherwise we use passive. This view of voice selectivity is contrary to the view of simple mutual paraphrasability. Furthermore, in Table 12, we noted that the frequency of causative cases is considerably (10 times) higher than passive cases. This fact suggests that the causative formation receives even stronger selectivity than passive. This higher selectivity leads the larger information entropy (which is a quantitative indicator of semantic markedness), and consequently will give a higher relevance estimation. Such explanation may be still too simplistic. For example, in Table 12, the frequency relationship between transitive and intransitive is quite different from the relationship between causative and passive, and we cannot explain why monovalent strategies degraded the performance. We need further investigation.

Next, in our experiments we were able to separate the detailed effects neither of the unaccusative and unergative (and their transitive partners), nor of various doublet-making morphological patterns, such as root categories (i.e., Vt, $\underline{V}$t, Vi, $\underline{V}$i) or ending patterns (i.e., type (A)-(F) in Table 3, including the short causative, etc.). As we discussed at length in section II.D-E, these two problems (i.e., unaccusativity and doublet morphology) are related to each other, and consequently will affect (grammatical) paraphrasability in certain ways and degrees. In addition, the existence of subject's agentivity distinctively characterizes the unergativity (versus unaccusativity), therefore this topic is related to the selectivity issue of the voice which was discussed in the previous paragraph.

In our experimental results, the superiority of the bivalent strategy over the monovalent one seems very clear, but the above simplistic explanation does not fully account for the phenomenon, and we need to ask further questions in a more careful analysis: why monovalent strategies did not work; or which (in)transitive domain or feature works, and which does not. In this valency dichotomy analysis, we have probably made one step forward, but have posed more research questions, as well.

3. Analysis of Linguistic Module Dichotomy (Lexicon vs. Syntax)

This performance analysis contrasts monovalent and bivalent query strategies, and corresponds to Hypothesis-LS [Lexical vs. Syntactical] (and two derived hypotheses, Hypothesis-TC [Transitive vs. Causative] limited by bivalency, and Hypothesis-IP [Intransitive vs. Passive] limited by monovalency).

The data here contrasted considerably with the result of the previous valency analysis. First, let us examine the results of strategically combined methods, i.e., lexical (=intransitive+transitive) and syntactical (=passive+ causative). While the effects of bivalent and monovalent approaches radically split into positive and negative sides, the difference in effectiveness between lexical and syntactical strategies is very small, both experiencing a similar level of positive improvement (~10%). This approximate equivalency is quite intriguing. (The lexical result was better than the syntactical, but the gap was insignificantly small.) Regardless of the conventional wisdom in information retrieval that less structural (i.e., lexical or morphological) handling (e.g., stemming, idiomatic noun phrases, etc.) should be the first choice as a promising method over syntactically structural methods, we are here claiming an equation of "lexical=syntactical" for retrieval. Thus, the combined effect of causative and passive is comparable to the combined effect of transitivity and intransitivity. (Note that we excluded the potential construction from the combined syntactic queries to make the comparison fair.) Although it occurred within our limited domain of verbal argument structures, and we are not arguing that such an equation has a universality in retrieval, it nevertheless encourages us to pursue grammatical and structural approaches to information retrieval further. We need more examination and exploration to get a broader and more solid picture of this research topography.

4. Valencies vs. Linguistic Modules

We have already observed a clear behavioral difference between valence data and linguistic module data. Thus, although valency change created a large performance gap between bivalency and monovalency, module difference between lexicon and syntax did not result in a significant difference. Thus, we claim significantly different profiles between the two dichotomies, so as to affirm Hypothesis-VL.

Based on this result, it is recommended to take either a lexical or syntactic valency control strategy in query processing. The advantage of the syntactic method is its higher applicability than the lexical method. The applicability of the lexical method depends on the productivity of the doublet making. The productivity of doublet making itself varies from language to language. Japanese is suitable for using this property in the query, but English is not. The advantage of the lexical method is a lesser dependency on structural analysis, which analysis is more costly in computation because local processing (such as processing of the permanent lexicon or morphological analysis) can be more responsible for the task.

Another recommendation is that if a query can be paraphrased with transitivization and/or causativization, applying the bivalent strategy may also boost precision. As we discussed for the theoretical model (II.E), the easiness or difficulty in causativizing the original query depends on the particular verb involved.

5. Verbs vs. Verbal Nouns

This is a question posed by Hypothesis-VVN [Verb vs. VN]. A noticeable performance difference between the verb and verbal noun results an inverse effect when operated on by the automatic method and the relevance feedback method. A possible explanation is as follows. Since verbal nouns generally have much less lexical ambiguity than verbs, the automatic method can choose the "right" verb-noun combination candidates with fewer mistakes. Thus, ambiguous terms produce "wrong" candidates among the verb-noun combinations that may appear somewhere in the text, so that the frequency data become distorted. However, verb and verbal noun data differ in their baselines, and we need more careful experimentation to confirm this result.

Another difference we may point out is the magnitude of improvement. Although many verbal strategies achieved around 10% improvement, most cases in verbal noun experiments showed around 6% or less improvement. One possible explanation is as follows. In the text, a considerable number of verbal noun expressions are realized in genitive constructions, and these tend to block causativization, passivization or potentialization. However, the genitive construction has a negative effect on retrieval, probably because of its structural inflexibility (which resists paraphrasing) and ambiguity (such as between subjective and objective genitives). Another possible reason may be found in the statistical characterization. For example, in Table 12 (section III.C), we found very large frequency gap between causative and passive for verbs (i.e., 0.8% vs. 8%, respectively), however it was relatively small for verbal nouns (i.e., 1.6% vs. 4.6%). Thus, verb causatives have much larger information amount than verbal noun causatives, and consequently, the experimental results of verbs likely show more distinctive characteristics than verbal nouns. Again, we need further experiments to confirm this result.

6. Relevance Feedback vs. Automatic

In III.D.6.b we introduced Hypothesis-RA [Relevance-Feedback vs. Automatic] concerning the performance difference between the relevance feedback method and the automatic method. Although this comparison has little relevance for linguistic justification, it is important in a practical area. We have already described the higher achievement of the automatic method for verbal nouns versus that of relevance feedback. For verbs, even though the results of the automatic method never exceeded those of relevance feedback, the difference became very close in the combined approaches. In any case, the results of the automatic method are encouraging, and we want to develop better statistical selection methods for verb-noun combinations than that which we tried in this study.

7. Single Technique vs. Combined Technique

Combinations of individual strategies (i.e., lexical as transitive+ intransitive, syntactical as causative+passive, and bivalent as transitive+ causative, except monovalent as intransitive+passive) seem to work comparably (in relevance feedback cases), or even better (in automatic cases) than any one technique used alone. However, they did not boost the performance as much as additively possible with individual improvements in contrast to some techniques which cumulate every single improvement. For example, the combination of character-based and word-based indexing led to better retrieval effectiveness by taking advantages of both methods (Fujii & Croft, 1994). The existence of such different effects of combinations probably depends on the characteristics of the particular combination of elementary methods. In our case, a combination consists of complementary elementary methods. For example, the lexical method consists of transitive and intransitive, and these elements are (nearly) mutually exclusive and complementary in a single classification framework. On the other hand, character-based and word-based methods are not complementary. In a complementary combination, the combined query does not increase, but rather eventually reduces the query's information (i.e., entropy), from the original elemental queries. Since our experiments used the #sum operator to create a combination that is not a purely additive operator (e.g., Boolean OR) nor multiplicative operator (e.g., Boolean AND), the system's behavior should be more complicated, but the above principle probably holds.

Nevertheless, fortunately, the degradation by our combination methods seems, practically speaking, insignificant.

There is another characteristic of the combined methods. In the data, the corresponding pairs of relevance feedback data and automatic data are highly correlated, and have less irregularity compared to the results of individual techniques. (The correlation coefficients are 0.98 and 0.12 for the combined methods and individual methods, respectively.) This smoothing effect can be explained by averaging the effects of the combined strategies. This property is a very desirable one for a practical system to have in order to obtain stable positive results.

D. Summary of Retrieval Experiments

In this chapter, we have conducted a series of experiments in order to observe various aspects of strategic linguistic applications on query formulation in retrieval. These experiments were mainly designed to fit in with a contingency of two dichotomies: the valency dichotomy and the linguistic module dichotomy. The valency dichotomy is based on the contrast between bivalency (i.e., the valency equals two, the subject and the object) and monovalency (i.e., the valency equals one, the subject only). The linguistic module dichotomy is laid out in the contrast between lexicon (i.e., transitive and intransitive) and syntax (i.e., causative and passive). Let us summarize the results of our experiments, aspect by aspect, according to the hypothesis.

1) **Valency control strategies improved retrieval performance**: Experimental results showed that valency control strategies (including potential, but excluding monovalent cases, i.e., intransitive, passive, or the combination of

these two), which is an example of grammatical paraphrasing, showed a considerable improvement in precision.

2) **Bivalent outperformed monovalent**: Bivalent experiments, i.e., transitive, causative, or a combination of these two, constantly achieved high improvement, no matter whether carried out through relevance feedback or the automatic method. The results showed not only a significant performance difference between them, but eventually a polarization between them, i.e., an improvement by bivalency, and a degradation by monovalency. But, some unique monovalent strategies, such as relevant feedback intransitive or automatic passive, still displayed a certain improvement in precision. Furthermore, there is still a chance one could improve the performance of monovalent strategies by applying a D-structural level analysis or other mechanisms. Thus, further experiments are necessary to determine the detailed effects of unaccusativity, suffix patterns, or related semantic characterizations. These are important not only for investigating the negative effects of monovalency, where unaccusativity clearly manifests itself, but also for getting a better picture of the mechanism of transitive counterparts, because these properties are associated with the construction of verbal doublets in morphology.

3) **Both lexical and syntactical methods performed positively, and comparably**: Both lexical and syntactic strategies in combined forms, i.e., transitive plus intransitive for lexical, and causative plus passive for syntactic, improved the precision. Also, their performances are close, even though the lexical method results were slightly better than the syntactic, in both relevance feedback and automatic methods. Furthermore, the improvements experienced by lexical and syntactic strategies are comparable to that experienced by the bivalent

strategy. Additionally, the potential strategy, which is syntactical, though it experienced no affect on the verb's valency in either relevance feedback or automatic, also showed a moderate improvement in precision.

4) **Verbal noun results were not as encouraging as those of verbal strategies; in particular, the genitive strategy degraded performance**: Retrieval performance by verbal noun strategies generally appears not as good as the results for verbs. These strategies seem to be more unreliable or unstable in achieving firm improvement. The best result for verbal noun construction was, despite the best result being obtained through bivalency in the case of verbs, achieved through the potential strategy, in both the relevance feedback and automatic techniques, though the reason is not clear. The query with the genitive strategy, which typically forms a noun phrase construction, showed a clear degradation in performance, probably a reflection of limited success of the phrase recognition in past IR studies.

5) **Sentence recognition improved performance**: It improved the precision significantly (at the highest 15% at the low recall level) from the performance of unstructured keyword query. This is not a part of valency control strategies, but a basic query technique based on sentence level language representation. This effect was found when the secondary sentence-level baseline for measuring the valency control effect was generated. In other words, the total precision gain from the keyword query to the best of our valency-controlled queries reached a considerable level (e.g., for verbs, +34.0% increase in the case of relevance feedback causative (averaged at 10% and 20% recall, as are the following figures); +30.6% in the case of relevance feedback bivalent; +21.2% increase in the case of automatic lexical case; +17.0% increase in the case of automatic transitive).

6) **Little stemming effect**: The stemming effect through transitivity normalization showed little improvement in performance.

7) **Automatic method performance was close to that of the relevance feedback method**: Especially when we combined more than one strategy (e.g., the "lexical" method consists of transitive and intransitive, etc.), automatic methods achieved performance results that were very close to those of relevance feedback. Thus, we obtained an encouraging result for the practicality of the grammatical paraphrasing approach.

8) **Combined methods produced more consistent and stable results**: The results of combined method appear more robust and consistent throughout the linguistic aspects (i.e., categories in Figure 21) than those obtained through a single strategy. The probable reason is the complemental relation of cues in the combination of strategies. Also as described in 7), a combination of strategies makes the performance of the automatic method close to that of relevance feedback.

According to the above observations, it is reasonable to recommend the following basic guideline in formulating a query: "First, if your query has tight verb-noun associations, insert them as a sentence pattern subquery into your query. Next, apply a lexical, syntactical, or bivalent (i.e., transitive, causative, or their combination) strategy on the above verb-noun combinations, then merge this strategic subquery into your query."

CHAPTER V

CONCLUSION AND FUTURE WORK

We started this dissertation by discussing what certain universal features of languages mean, and fitting them into a model of information retrieval, namely in the perspective of the language evidence paradigm. The inference network model fit well in this framework as we later saw in our methodology and experiments. In those discussions, we used well-defined linguistic conceptual constructs as a background framework to provide accountability, but these models still need the affirmation by experimentation with certain techniques, which may lead to a workable "linguistic" retrieval system. This modeling-experimentation process should be regarded as a part of the theory-practice cycle, which has often been overlooked in the empirical strains of IR system research. Nevertheless, "natural language processing" has not received full appreciation as a maturely accountable technology for information retrieval, and the term "linguistic information retrieval" itself is even not an established technical terminology in IR research. Thus, what is most vital is to show proof of the existence of an applicable and effective domain of linguistic retrieval, and this dissertation has, we believe, served this role through the following methodology and experimentation.

What we chose as our conceptual device was called the grammatical paraphrasing model. It was a paraphrasing (in our case, of a query) operation, but was not a semantic-oriented arbitrary manipulation, but rather a formally definable grammatical operation. We proposed a valency control strategy, which controls the valency of a given verb in a sentence by lexical or syntactic means, as a sample realization of the grammatical paraphrasing model. In the framework of valency control strategies, two dichotomies are laid out, namely the valency dichotomy and the linguistic module dichotomy. The former concerns two opposing states: bivalency (represented by transitive and causative) and monovalency (represented by intransitive and passive). The latter is the contrast between lexicon (represented by transitive and intransitive) and syntax (represented by causative and passive). These dichotomies pose very interesting and valuable research questions concerning disparity of retrieval performance. The first question is as follows: Which is a better retrieval device to improve effectiveness: transitive or intransitive? causative or passive? or, in general, bivalency or monovalency? The second question is as follows: Which is a better retrieval device to improve effectiveness: transitive or causative? intransitive or passive? or, in general, lexical or syntactical? The Japanese language provides an excellent test-bed for examining these questions because the Japanese verb system, in contrast to that of English, exhibits extensive transitive-intransitive doublets with their rich verbal morphology, so that the effects of lexical valency and syntactic valency are testable in a parallel manner. Additionally, the Japanese morphology makes the indexing task easy in terms of marking the various linguistic features necessary for our experiments, such as subject, object, transitive, intransitive, causative, passive, etc. In addition, we were also able to look at the effects of potential and verbal nouns including the genitive formation.

Our observations from our experiments were extensively portrayed in the summary of experiments (section IV.D). Here, let us repeat our four most important findings:

1. Retrieval performance, especially precision was improved by the use of valency control query strategies, except by monovalent kinds. Thus, the grammatical paraphrasing paradigm as a linguistic retrieval methodology works.

2. Bivalent outperformed monovalent. In other words, transitive is better than intransitive, and causative is better than passive. However, since intransitivity involves a finer mechanism, which is manifested by unaccusativity or related morphological patterns, we might have further chances to develop better strategies even for monovalent situations.

3. Both lexical and syntactical methods performed positively, and comparably. In other words, retrieval with transitive and causative, or intransitive and passive, are comparable.

4 For verbal nouns, especially in genitive constructions, results are not as encouraging as those for verbs.

In addition to the above, from the view of practicality, we add the following:

5. The automatic method performed close to the relevance feedback method.

Finally, to close this dissertation, let us describe the possible directions of our future research. First of all, we need larger scale experiments, especially with more convincing numbers of queries, to confirm the experimental results of this dissertation. With sufficient queries, we will not only have more reliable data, but also will be able to determine in more detail, the retrieval effects of unaccusativity (or unergativity) and/or transitive (or intransitive) suffix patterns, which include short causative (or short passive) and others, as both may be manifested in association with the lexical-syntax dichotomy with certain semantic characterizations. This exploration might lead to better handling of monovalent queries. In the same line of analysis, we may want to investigate the nature of distinction between doublet and non-doublet verb types (which varies with the specific language as seen at (7) and (8) in section II.C). If retrieval factors of both doublet and non-doublet verbs are successfully characterized, the applicable domain of valency control strategies will be considerably expanded. To improve the performance of verbal noun strategies, we may also want to look at similar grammatical relationships within a (nominal or verbal) compound (e.g., see Kageyama, 1993).

Another way to expand our approach is to apply our strategic query formulation paradigm to other languages. Our next target can be English, because we already have basic data on the performance comparison between English and Japanese (Fujii & Croft, 1994). Other Asian languages such as Chinese or Korean, or other non-English Western languages such as Spanish, French or German may be also interesting targets. If this could lead to connections between the retrieval behavior of an IR system and some language parametric or typological feature, the horizon of universal applicability in retrieval will be greatly expanded. Note that if we want to achieve such an ultimate expansion, a more solid linguistic accountability will be required in our retrieval system. In this dissertation, we successfully demonstrated that the inference network model could take advantage of linguistic evidences as knowledge sources. Its post-coordinative operators made it possible to flexibly tailor strategies in actual query formulae. However, more language-oriented powerful operators/functions, such as parsing, pattern matching, etc., are desirable.

Lastly, we may want to expand our scope of grammatical technique of information retrieval into a larger linguistic arena. Such a possible challenging research area is the aspect of *agentivity*. We briefly touched this notion in the analysis of valency dichotomy (section IV.C.2). For example, in a valency control strategy, the choice of passive sentence was addressed according to the agent selectivity. Thus, the issue of agentivity is related to both grammar for competence (e.g., passivization) and language performance (e.g., voice choice). In addition, the agentivity property also associates with certain typological features such as subject-prominence versus topic-prominence (Li and Thompson, 1976). A carefully integrated approach of linguistically distinct dimensions must be developed for the future of linguistic information retrieval.

BIBLIOGRAPHY

Attar, R., Choueka, Y., Dershowitz, N., & Fraenkel, S. (1978). KEDMA: Linguistic tools for retrieval systems, *Journal of the Association for Computing Machinery, 25* (1), pp. 52-66.

Baker, M. (1988). *Incorporation: A theory of grammatical function changing*. Chicago, IL: University of Chicago Press.

Ballesteros, L., & Croft, W. B. (1996). Statistical methods for cross-lingual information retrieval, *Workshop on Cross-Lingual Information Retrieval,* Zurich, Switzerland.

Callan, J. P., Croft, W. B., & Harding, S. M. (1992). The INQUERY retrieval system, *3rd International Conference on Database and Expert Systems Applications*.

Carstairs-McCarthy, A. (1992). *Current morphology*. New York: Routledge.

Choueka, Y. (1990). Responsa: An operational full-text retrieval system with linguistic components for large corpora, *Proceedings of LALL, also in Computational lexicology and lexicography (A. Zampoli (Ed.), 1990, p. 150).*

Croft, W. B. & Xu, J. (1994). Corpus-specific stemming using word form co-occurrence, *Proceedings of the 4th Annual Symposium on Document Analysis Information Retrieval*. Las Vegas, Nevada.

Croft, W. B., Broglio, J. & Fujii, H. (1995). Applications of Multilingual Text Retrieval.

Croft, W. B., Turtle, H. R., & Lewis, D. D. (1991). The use of phrases and structured queries in information retrieval, *ACM SIGIR-91*, pp. 32-45.

Davis, M. (1996). New experiments in cross-language text retrieval at NMSU's computing research lab. In D. K. Harman (Ed.), *The Fifth Text REtrieval Conference (TREC-5)*. NIST.

Dillon, M. & Gray, A. S. (1983). FASIT: A fully automatic syntactically based indexing system, *Journal of the American Society for Information Science, 34*, pp. 99-108.

Fagan, J. (1987). Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and non-syntactic methods, *Ph. D. dissertation, Cornell University*.

Frakes, W. B. & Baeza-Yates, R. (1992). *Information retrieval: Data structure and algorithms*. Englewood Cliffs, NJ: Prentice Hall.

Fujii, H., & Croft, W. B. (1993). A comparison of indexing techniques for Japanese text retrieval, *Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp.237-246)*. Pittsburgh, PA.

Fujii, H., & Croft, W. B. (1994). Comparing the retrieval performance of English and Japanese text databases, *Proceedings of the 2nd Annual Workshop on Very Large Corpora*, pp. 87-97.

Fujii, H., & Kitagawa, C. (1997). Transitivity Alternations in Japanese. Manuscript submitted for publication.

Fung, P., & Wu, D. (1994). Statistical augmentation of a Chinese machine-readable dictionary, *Proceedings of the 2nd Annual Workshop on Very Large Corpora*, pp.69-85.

Gachot, D. A., Lange, L., & Yang, J. (1996). The SYSTRAN NLP browser: An approach of machine translation technology in multilingual information retrieval, *Workshop on Cross-Lingual Information Retrieval, SIGIR '96*.

Gass, S. M. (1989). Language universals and second language acquisition, *Language Learning, 39* (4), pp. 497-534.

Gay, L. & Croft, W. B. (1990). Interpreting nominal compounds for information retrieval, *Information Processing and Management, 26* (1), pp. 21-38.

Harada, T., et al. (1989). Reduction of search noise by using role indicators, *JICST Procs. of the 26th Annual Meeting on Information Science and Technology*, pp. 139-144.

Harman, D. (1992). The DARPA TIPSTER Project, *SIGIR Forum, 26(2),* pp. 26-28.

Hayatsu, E. (1989). Yuutsui Tadooshi to Mutsui Tadooshi no Chigai-ni tsuite: Imiteki-na Tokuchoo-o Chuushin-ni [On the Semantic Difference between Paired and Unpaired Transitive Verbs in Japanese], *Gengo Kenkyuu*, 95, pp. 231-256.

Hull, A. D., & Grefenstette, G. (1996). Querying Across Languages: A dictionary-based approach to multilingual information retrieval, *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 49-57.

Inoue, H. (1984). *SikaBan NihonGo-Bunpou (Japanese grammar, a private view)*. Tokyo: Shinchou.

Jacobsen, W. M. (1992) *The Transitive Structure of Events in Japanese*. Tokyo: Kuroshio Shuppan.

Kageyama, T. (1993). *Bunpou-to go-keisei (Grammar and word formation)*. Tokyo: Hitsuji Shobou.

Kageyama, T. & Shibatani, M. (1989). *MojûruBunpou-no GokeiseiRon: NO-Meishiku-karano FukugouGoKeisei (Word Formation in Modular Grammar: Compound Formation with "NO" Noun Phrase)*. Tokyo: Kuroshio.

Krovetz, R. (1995). *Word sense disambiguation for large text databases*. Doctoral dissertation, University of Massachusetts.

Lee, J. H., & Ahn, J. S. (1996). Using n-grams for Korean text retrieval, *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp.216-2224)*. Zurich, Switzerland.

Levin B. & Rappaport Hovav, M., (1995). *Unaccusativity: At the syntax-lexical semantic interface,* Cambridge, MA: MIT Press.

Li, C. N., & Thompson, S. A. (1976). Subject and topic: A new typology of languages. In C. N. Li (Ed.), *Subject and topic*. New York: Academic Press, pp. 457-489.

LINGUIST List, (1996). Vol-7-1005, July 10, 1996.

Marantz, A. P. (1984) *On the Nature of Grammatical Relations*, MIT Press, Cambridge, MA.

Matsumoto, Y., et al. (1991). User's guide for the JUMAN system: A user-extensible morphological analyzer for Japanese. Nagao Laboratory, Kyoto University.

McCawley, J. D. (1978). Notes on Japanese clothing verbs, in Problems in In John Hinds and Irwin Howard (Eds.), *Japanese Syntax and Semantics*, pp. 68-77.

Mikami, A. (1960). *Zou-wa hana-ga nagai (Concerning elephants, their noses are long).* Tokyo: Kuroshio.

Miyagawa, S. (1989). *Structure and Case Marking in Japanese* (*Syntax and Semantics* 22), Academic Press, New York, NY.

Nie, J. Y., Brisebois, M., & Ren, X. (1996). On Chinese retrieval, *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp.225-234).* Zurich, Switzerland.

Perlmutter, D. (1978). Impersonal passive and the Unaccusative Hypothesis, *Proceedings of the Fourth Annual Meeting of the Berkeley Linguistics Society*, Berkeley Linguistics Society, University of California, Berkeley, pp-157-189.

Perlmutter, D. & Postal, P. (1984). The 1-Advancement Exclusiveness Law. In D. Perlmutter and C. Rosen (Eds.), *Studies in relational grammar, 2*, pp. 81-125.

Pevzner, B. R. (1972). Comparative estimation of the operation of the Russian and English modification of the Empty-Nonempty-2 system (in Russian), *Nauchno-tekhnicheskaya Informatsiya, Seriya 2, 6,* pp. 31-33.

Porter, M. (1980). An algorithm for suffix stripping, *Program, 14* (3), pp. 130-137.

Qiu, Y., & Frei, H. P. (1993). Concept based query expansion, *Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp.160-169).* Pittsburgh, PA.

Roeper T. & Siegel, D. (1978). Lexical transformation for verbal compounds, *Linguistic Inquiry, 9*, pp. 199-260.

Salton, G. (1971). Automatic processing of foreign language documents, *The SMART retrieval system: Experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice Hall.

Salton, G. (1983). Introduction to modern information retrieval, *Information Processing Letters*. NY: McGrow-Hill.

Selkirk, E. (1982). *The syntax of words (Linguistic Inquiry Monograph, 7)*. Cambridge, MA: MIT Press.

Shibatani, M. (1973). Semantics of Japanese Causativization, *Foundations of Language* 9. 327-373.

Shibatani, M. (1976). Causativization, in Masayoshi Shibatani (ed.), *Japanese Generative Grammar* (*Syntax and Semantics* 5), New York, NY: Academic Press, pp. 239-294,.

Shibatani, M. & Kageyama, T. (1988). Word formation in a modular theory of grammar: Postsyntactic compounds in Japanese, *Language, 64*, pp. 451-484.

Sparck-Johns, K. (1972). A statistical interpretation of term specificity and its application in retrieval, *J. Documentation, 28* (1), pp. 11-20.

Spencer, A. (1991). *Morphological theory*. Oxford, UK: Blackwell.

Svartvik, J. (1966). *On voice in the English verb*. Hargue, Netherlands: Mouton.

Tanaka, H. (1989). *ShizenGengoKaiseki-no Kiso [Foundations of Natural Language Processing]*, Tokyo: Sangyo Tosho.

Turtle, H. (1991). Inference networks for document retrieval, *Ph. D. dissertation, University of Massachusetts*.

Turtle, H. & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model, *ACM Transactions on Information Systems, 9* (3), pp. 187-222.

Van Rijsbergen, C. J. (1979). *Information Retrieval (2nd Ed.)*. Boston, MA: Butterworth.

Voorhees, E. M. (1983). Using WordNet to disambiguate word senses for text retrieval, *Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp.171-180).* Pittsburgh, PA.

Wien, C. (1996). A new methodology for testing retrieval efficiency for multiscriptual databases, applied on testing the efficiency of Arabic natural language in OPACs, *Workshop on Cross Language Information Retrieval, Zurich, Switzerland,* pp. 66-98.

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis, *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 4-11.