American Journal of Computational Linguistics *Microfiche 48*

ABSTRACTS OF THE 1976 INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS -COLING-

MARTIN KAY
Program Committee Chairman
Xerox Palo Alto Research Center
3180 Porter Drive
Palo Alto, California 94304

EDITORIAL NOTE

The abstracts published here were prepared from the longer summaries submitted by contributors. The selection was made by the Program Committee for ICCL 76 The staff of AJCL must accept responsibility for any distortion that occurred in reducing the length of each summary to fit on a single microfiche.

ICCL will take place June 28 - July 2, 1976. Information can be obtained from

COLING 76, Department of Linguistics
University of Ottawa
Ontario, Canada K1N 6N5

An alphabetical list of contributors, with addresses as supplied, begins on frame 81 of this fiche.

--DGH

CONTENTS

SESSION I

SESSION II:

SESSION III:

SESSION IV:

SESSION V:

SESSION VI:

SESSION VII:

SESSION I

Plenary Lecture

THE NEED FOR A FRAME SEMANTICS IN LINGUISTICS

Charles Fillmore

ON THE NOTION OF SEMANTIC LANGUAGE

PETR SGALL

1. Many models of natural language understanding, man-machine communication, etc. are being constructed, and many semantic or cognitive languages are being proposed to serve as input and output languages of the brain of such systems. Some requirements on such languages have been formulated; for instance they should include no ambiguities, they should allow for an effective and empirically adequate deductive or inference procedure, they should be as close to natural language as possible (to permit an economical analysis and synthesis of natural language input and output for the whole system), and at the same time they should allow for a relatively easy implementation in computers. However, a systematic discussion of such requirements that would ensure such a language to deserve its attribute "semantic" is still lacking. From a strictly logical viewpoint the proposed languages, most of which are not fully formalized, are often regardes as ad hoc

2. However, an examination of the tools logic offers for explicit semantics shows that some of the crucial problems of natural language are still unsolved. Even if such devices as modal and tense logic, possible worlds, pragmatical indices (or points of reference) and intensional meanings are used, the meanings of two (analytic or synthetic) sentences with identical truth conditions cannot be held apart.

3. A purely formal treatment of the "sense" of sentences appears to be excluded; there seems to be a single possibility how to account for the "sense" of sentences in an explicit way, viz. to formulate a procedure translating the (deep or tectogrammatical structure of the) sentence into a language which, in the end, must be, of course, interpreted without

any formalism. It appears that this final interpretation must be simple enough to make any further formalization superfluous, i.e., the semantic language must be transparent, it must have a simple syntax, each of its rules having a single interpretation with regard to a given model.

[redacted] thus can be concluded that a systematic discussion of the re[redacted]uirements of a semantic language is necessary for practically oriented man-machine communication research as well as for the theory of cognitive psychology and of logic

WHAT'S IN A CONCEPT: STRUCTURAL FOUNDATIONS FOR SEMANTIC NETWORKS

RONALD J. BRACHMAN

I wish to deal with some fundamentals of semantic networks and make explicit some assumptions in order to get a better grasp of representational power and develop a criterion of "well-definedness" Network notation is "associative"; Woods' analysis of "what's in a link" points out that the standard repertoire of links is insufficient. Here we will take a look at what we want a node to represent and how to represent it in a consistent and well-defined way.

To an almost universal extent, a node is used to represent a particular object or event or a class of objects or events. "ISA", etc., links implicitly express more than class membership; by virtue of a link, one node has all of the properties known to be attributable to the other. Hence the network formalism must allow for the representation of the properties of a class of objects.

We introduce a new primitive link type called DATTRS ("define as attributive parts" The description is a node which expresses, among other things, the ROLE that a part plays in making up the object, and the set of values that the part can have. The part description is a node; at it we can express an arbitrarily complex description. It is not necessary to discriminate between physical and other kinds of attributes. By separating the ROLE and the VALUE/RESTRICTION, we can modify the description of a DATTR without compromising its functional role in the whole concept. Sets of independent assertions are insufficient to discriminate between objects with identical parts but unsimilar arrangements; STRUCTURAL/CONDITION indicates relations. We apply this theory to the problem of assimilation of new information

TEXT-BASED LEXICOGRAPHY AND ALGORITHMIC TEXT ANALYSIS

STURE ALLEN

When a linguistic model is applied to a large corpus of authentic text, a considerable number of problems inevitably present themselves. Works of reference naturally tend to disregard many of these things. Studies of special issues, furthermore, often presuppose that the solutions needed in the particular case are in fact at hand, even if they are not (which is, as a matter of principle, of course justifiable). In any event, this provdes two reasons for a close study of a large corpus on an explicit theoretical basis Needless to say, there are other reasons, too, in particular the intrinsic descriptive interest of such an overall analysis and the possibility of gaining new insights into the nature of language.

In this paper, some aspects of the co-ordinated work of three groups in the Department of Computational Linguistics at Göteborg University are reported. The first is the Research Group for Modern Swedish, which carries out the work underlying the volumes of the Frequency Dictionary of Present Day Swedish. The second is the group developing the Swedish Logotheque, a text and word bank. The third is the group working on the project Algorithmic Text Analysis.

The dictionary project aims at a quantitative and qualitative lexical analysis (in a wide sense) of a corpus of one million running words from five morning papers, representing Present-Day Standard Swedish in its written form. The general strategy developed for the investigation is presented. Comments are made on the significance of the constructional tendency of words, the role of discontinuous collocations, generation of coherent text, stylistics, and psycholinguistics.

THE USE OF WORD-CLASS DISTRIBUTION DATA FOR STYLISTICS

DONALD ROSS, JR.

I am developing a taxonomy of structural properties of literary texts in order to describe similarities and differences among the styles of various authors A text has many features, some well represented by standard statistical measures, some not Current approaches can not explain how features from various levels interact or how to represent dynamic changes within a text What can be done is to build on descriptions of structural characteristics that can serve as the basis for comparing indices and developing a coherent view of text structure This study develops statistical indices for large samples from various authors. It uses an integrated sequence of programs, named EYEBALL, to count, parse, and analyze descriptive statistics I have analyzed samples of English Romantic poetry, and other users have worked on samples of novels

I propose to compute a series of measures that maintains the known linguistic realtions without losing any of the initial data. A list of categories is arranged under five headings nominal-phrase constituents, verbal-phrase constituents, adverbs, conjunctions, and miscellaneous Linguistic relations (phrase-head choice and modifier-head ratio) are defined. The procedure computes twelve measures, their averages variances, correlation coefficients, and linear regressions

Data from Keats, Blake, and Coleridge, and dialogue and narration in Joyce's Ulysses, illustrate the approach The small size and heterogeneity of the base prevent our making conclusive generalizations, but once the base is large enough we can provide "background" or "norms" for genres and eras, and evaluate differences and similarities

A COMPILING SYSTEM FOR AUGMENTED TRANSITION NETWORKS

RICHARD R. BURTON AND WILLIAM A. WOODS

The ATN formalism was developed as a representation for natural language grammars. The grammar has been viewed as a data structure which is interpreted by a program (parser). This paper describes a system that views the ATN as a machine description that can be compiled into a runnable program. The significance is a dramatic reduction in processing time. The compiled programs parse ten times faster than LUNAR.

The ATN is a description of "what sentences the machine should accept"; the compiled ATN is additionally a description of "how the machine should accept them". The compiler decides about characteristics of the process left unspecified by the formalism. The first decision is what constitutes a configuration of the ATN machine (the amount of information needed to characterize its status) The second decision is what control structure the ATN machine is to have (order of trial of alternative parsing options). The compiler must decide how each arc is to be compiled into code which, when executed, will change the configuration as desired; choices range from how the nondeterministic process is to be translated into a deterministic program to choice of storage structures.

The compiler allows the user to choose a subset of features of the ATN formalism, and takes advantage of these choices to optimize ATN parsers. The compiler is limited to a depth-first control structure, but implementation of other strategies is possible and several are planned. The paper describes the system and presents some trade-offs which were explored. An example of operation is included. The possibilities of producing ATN machines in languages other than LISP are discussed. Efficiency has been tested on the LUNAR grammar and the SOPHIE semantic grammar.

ON A SYNTAX ANALYSIS METHOD FOR DEPENDENCY GRAMMARS IN A SPEECH RECOGNITION SYSTEM

SHUN-ICHI TAKEYA

The importance of linguistic information for successful speech recognition has long been acknowledged. In most systems, however, word segmentation is performed sequentially from the top of the input utterance (i.e. left-to-right), and use of syntactic information is mainly intended to constrain which words are to be matched with input. Such a scheme often comes to a deadlock, when it encounters an extremely indistinct part in an utterance. This paper proposes a system in which a syntax analyzer uses a top-down, breadth-first strategy.

We use a dependency grammar to describe syntactic structures. Dependency grammars were introduced by Hays and, in comparison with CG grammars, (1) can represent directly relations between terminals, (2) nonterminals correspond to word classes, (3) every expansion of a nonterminal produces a word corresponding to it. These features, especially (2) and (3), are very convenient for speech recognition. Namely, using a top-down parsing, we can analyze input by catching words successively.

The physical inputs for our system are utterances of the sentences which are generated by a dependency grammar. The first part of the system processes an input utterance according to its physical features; the output is an incomplete phoneme sequence Q. The second part processes Q under the control of part 3 and brings out a word sequence: Part 3 chooses an initial nonterminal and continues to expand nonterminals. At each expansion, part 3 attempts word matches in Q. Bactracking can occur. When Q is decided to have a suitable structure as a sentence, part 4 changes it into a letter sequence for output

In three experiments on 20 utterances in Japanese, our system proved effective.

UTILISATION DES REDONDANCES POUR L'ANALYSE ET LE CONTROLE AUTOMATIQUES D'ENONCES EN LANGUE NATURELLE

JACQUES COURTIN

Les systèmes d'analyse des langues naturelles reposent en général sur les deux grands principes suivants: le texte d'entree est correct, l'utilisation des divers modules est hiérarchisee. Il s'ensuit que les systèmes proposés répondent par oui ou non et délivrent dans le cas d'une reponse affirmative certains renseignements d'ordre linguistique utiles au modèle suivant. L'enchainement habituel des modèles est la succession "morphologie - syntaxe - sémantique." Ils sont par consequent incapables de localiser les erreurs et encore moins de proposer une ou plusieurs corrections possibles.
La mise au point des modeles linguistiques est difficile bien qu'il y ait séparation entre les parametres linguistiques et les programmes. L'apparition des systèmes conversationnels de programmation a donné naissance à des programmes plus ou moins sophistiqués d'analyse des langues. Le travail du linguiste s'en est trouvé facilité bien qu'actuellement les systèmes existants se contentent de proposer un fonctionnement conversationnel peu interactif a partir d'un terminal D'autre part, les modèles informatiques ne peuvent etre qu'une approximation des phénoménes linguistiques. Il s'ensuit qu'ils ne sont jamais complets, et qu'ils contiennent toujours des erreurs: il faut sans cesse y apporter de nombreuse retouches. On doit donc essayer de donner aux linguistes un outil assez souple qui leur permette de vérifier facilement l'adéquation du modèle proposé. L'idéal serait de proposer un systeme d'analyse permettant la localisation des erreur liees à l'insuffisance des modèles et une interaction totale entre l'utilisateur et la machine. C'est pour remedier a toutes les difficultés evoquées précédemment que le systeme P.I.A.F. a ete cree

ABERRANT FREQUENCY WORDS AS A BASIS FOR CLUSTERING THE WORKS IN A CORPUS

ALASTAIR McKINNON

This paper describes a method for clustering the works in a corpus by matching the aberrant frequency words in each work with those in every other work. Words are said to have an aberrant frequency if they show $\geq$ 1.96 standard deviations from the corpus norm. Given the list of such words for each work, the similarity index for each pair of works is then computed using the formula

$$\frac{C \times 100}{A + B - C}$$

where A is the number of aberrant frequency words present in the first work but not in the second

B is the number of aberrant frequency words present in the second work but not in the first

C is the number of aberrant frequency words common to both

The resulting similarity indices are then used as input to a multi-dimensional scaling programem to give an overall clustering of the works.

The paper includes a discussion of the method and its presuppositions and briefly assesses the results.

SESSION II

PLENARY LECTURE

TOWARDS A MODEL OF LANGUAGE PRODUCTION: LINGUISTIC AND COMPUTATIONAL FORMALISMS

HENRY THOMPSON

FRAMES, STORIES, SCRIPTS AND FANTASIES

YORICK WILKS

In Minsky's original paper, one can distinguish static and dynamic notions of frome: fixed situation or scene vs. sequence of events. I look at the latter notion and question whether the current explications are the sort, or level, of knowledge necessary for natural language understanding. I use the example of a puberty rite from a remote culture and argue that we can understand it without reference to the frame which we most likely do not have for it.

The thesis behind the application of the dynamic frame seems to be: 'In order to understand a story we need to know how basic stories of that type go' I call this the plot line hypothesis; it is supported by appeal to Bartlett's work on memory. But no evidence has been produced that a computational discourse understander needs such a thesis to function. I conclude that (a) some far more general inference rules might well do the trick, (b) plot line frames do not solve the 'topic' problem.

A strong version of the plot-line hypothesis hints at a vicious regress. It is not clear what mechanisms of access to frames will allow them to solve the "topic problem". Unless systematically related to smaller-scale mechanisms frames may be no advance on the older thesaurus hypothesis as far as topic is concerned. The paper also examines "do-it frames".

Advocates of frames sometimes implicitly argue from the true premises that (a) we need representations of knowledge to understand language and (b) we know (i) how to do certain things and (ii) how stories usually go on to the conclusion (c) we need representations of (i) and (ii) to understand language. The conclusion is not proved and, I suspect, false.

SEMANTIC FRAMES, SEMANTIC FIELDS, AND TEXT ANALYSIS

HEINZ J. WEBER

This paper deals with the question, how to build up meaning representations of natural lexical units (especially verbs and nouns, with regard to their semantic and syntactic roles in isolated sentences, and with regard to the various actualizations of properties or of actions and action participants in texts. This work is to be seen in connection with a parsing system for German and Russian; an identificational grammar cooperates with a lexicon. Text analysis consists in gathering sentences by interpreting identified sentential actions as belonging to more complex actions and by discovering coreferent action participants.

The lexicon is the main subject of this paper. In it, information includes semantic and syntactic frames, several semantic frames can be linked together within one synactic frame. The method discussed here is in the main directed by the goals that the program is intended to achieve; there is not claim to exhaustiveness or primitivity in a linguistic or cognitive sense. The labeling of nouns constitutes lexical classes, which lie between the traditional "Wortfelder" and the Fillmorian "deep cases", being more bound to sentence relations than the former, and more static than the latter, because their meaning is restricted not only by roles in semantic frames but also by features like SOLID, WEIGHT, SHAPE. Confi gurations of features are a special type of lexical unit, or "prototype word". Such units are codified in a "prototype" lexicon, which is reserved for text analysis, and so need not be identified in the sentence-analysis lexicon.

SUR LA CONFECTION D'UN LEXIQUE POUR L'ANALYSE AUTOMATIQUE

MORRIS SALKOFF

Le problème de la classification des mots dans un lexique faisant partie d'un analyseur automatique de phrases represente un des problèmes les plus difficiles pour un programme d'analyse automatique. Une classification sommaire des mots en un petit nombre de catégories principales n'est pas valable, car elle aboutit à un foisonnement d'analyses linguistiquement injustifiables ou sémantiquement incohérentes lors de l'analyse d'une phrase par ordinateur. La classification automatique des mots, realisee par programme sur ordinateur, ne se revele pas rentable non plus, et aboutit aussi soit à une perte de temps, soit à la production d'analyses incohérentes. La constitution d'un lexique 'automatique' doit etre basée sur des principes et des phenomènes linguistique qui sont en grade partie formalisables ceci exige un travail minutieux et long, mais a pour resultat l'elimination de beaucoup d'analyses incorrectes, souvent appelées 'parasites'

Un genre de dialogue entre le linguiste et l'ordinateur pourrait etre instaure. En analysant des textes, le programme fournirait de temps en temps un analyse injustifiable ou incohérente. Quant le linguiste examine cette analyse de plus près, il decouvre une nouvelle sous-classe exprimant un phénomène linguistique jusqu'alors insoupconne. L'analyse automatique peut donc servir comme processus de découverte de nouvelles sous-classes qui seront utile dans l'analyse effectuée ultérieurement.

MACHINE DICTIONARY AND LEXICON

G. FERRARI AND I PRODANOF

The MD, i.e. a list of lemmatized forms of Italian, was conceived both as an instrument in the procedure of lemmatization and as the nucleus for research. This initial precept has demanded, inter alia, our complete neutrality in relation to all thoeries. The list of entries has been defined a priori as the largest possible. All forms have been generated; archaic, popular, and rare ones included. Every record was provided with the information defining both some relationships between lemmas and different linguistic subsets to which they belong. An updating procedure, adding missing words and new information, has been set up.

The creation of computational models implies a change of our original theoretical neutrality. From this point of view, the MD ceases to be a simple list of lexical units and must be considered as a machine composed of a nucleus of data and a series of procedures working various linguistic levels.

The dictionary is situated between lexicology, morphology syntax and semantics. Between the latter two, the boundary is not precise, but from the theoretical and practical views, no refined analysis is possible without using a dictionary. It is particularly important to have a partially autonomous idea of the dictionary as a unitary image of lexical competence. In practical terms it is possible to associate with each lexical unit information which represents lexical competence and also may be interpreted algorithmically and converted into specific codes for each level of analysis.

VERS UN MODELE ALGORITHMIQUE POUR LE TRAITEMENT AUTOMATIQUE DES LANGUES NATURELLES

J. CHAUCHÉ

Le traitement automatique des langues naturelles nécessite deux theories: une théorie linguistique et une théorie algorithmique. Une interaction entre les deux est nécessaire La théorie linguistique doit permettre une analyse satisfaisante, mais elle doit etre degagée de toute algorithmique. La théorie des graphes étiquetés parait adaptée à la fois pour la représentation d'une analyse relationnelle et d'une algorithmique. La notion d'arborescence permet l'élaboration d'algorithme de manipulation où la description des transformations devient statique et n'est pas basée sur le cheminement. La notion de graphe multi-étiqueté et la notion d'étiquette référence permet à la fois de concentrer la représentation et d'obtenir un graphe très proche d'une arborescence ou la liaison entre certain points indépendants suivant la relation principale est réalisée. La présentation de ce modèle est divisée en trois parties: les objets manipulés, la façon de les manipuler par des transformation élémentaires et enfin la définition de grammaires générales de manipulations. Le modèle défini ice est une extension d'un modèle expérimental, le système CETA. L'étude de ce système montre la souplesse d'utilisation d'un tel modèle pour l'analyse des langues naturelles et l'efficacité importante due a l'existence d'algorithmes puissants pour la manipulation d'arborescences. Les extensions definies ont pour but de palier aux inconvénients rencontrés par les linguistes dans la définition des grammaires d'analyse et de synthèse. La programmation en cours de ce modèle est faite dans un langage de haut niveau, ce qui permettra une meilleure transportabilité.

A FRAMEWORK FOR LANGUAGE UNDERSTANDING

WILLIAM H. PAXTON

This paper describes a framework for a unified approach to four major problems: Integration of contributions and interactions of multiple knowledge sources; cooperation; evaluation; and attention to avoid thrashing among a number of competing alternatives. These problems result from entering the large space of possibilities in a language with error prone, imprecise knowledge. In the attempt to develop speech understanding for a substantial subset of natural English these problems have been unavoidable.

We choose the phrase as the basis for integration; it is the natural unit of structure and meaning. A parse net brings together all attempts to construct a particular category of phrase starting or ending at a particular location in the input. To reduce the cost of evaluation, heuristic methods are used to search the parse net for a best path. To shift attention when the first selection leads to poor results, our system first puts words in focus when an incomplete phrase including them is selected for processing, and then inhibits work on other phrases that are inconsistent with the focus, by lowering the priority of tasks. If a phrase in conflict with focus overcomes the bias against it. the system's attention shifts to a new focus.

Illustrations of the use of this framework for language understanding are drawn from the speech understanding system being developed jointly by SRI and SDC.

SEMANTICS IN AID OF AUTOMATIC TRANSLATION

THOMAS R. HOFMANN

The naive incorporation of an integrative representation as a 'pivot language' is impossible in the near future and, in any case, undesirable; re-expression leads not to a translation but to a version that does not employ synonymous lexical items, parallel syntactic structures, or even the same order of expression. It is sometimes important that translation is a "maximal isomorphism which preserves the semantic meaning"

A semantic representation can be used to monitor output and to redirect translation when needed to maintain equivalence

One way to incorporate semantics into an AT system is to arrange for structural transfer at the lowest level possible, thus employing a minimum of analysis of the source language and of synthesis of the target language. The TL sentence is then analyzed to determine if it has the same semantic effect as the SL sentence which was analyzed first.

Another way is to make a complete syntactic analysis of the SL sentence at one go, but to use only selected aspects of that analysis for TL synthesis. Again, semantic analysis determines if the TL sentence is equivalent.

There appears to be nother way to obtain what we normally refer to (and expect or even require) as translations. I will compare these two solutions to show that they are only engineering varieties of the same general solution, which preserves meaning at the cost of differences in structure, but so far as possible preserves structure as well.

ORGANIZING KNOWLEDGE FOR ENGLISH-CHINESE TRANSLATION

WALTER J. STUTZMAN

This paper discusses an English-Mandarin translation system knowledgeable about two domains: trips and restaurants. The system has four components: an English parser, a Mandarin generator, the Script Applier Mechanism SAM, and a memory The main problem in translation, as in other language generation tasks, is "what to say". This problem includes not only the production of well-formed sentences but the more complicated problem of organizing and using knowledge to produce reasonable explanations for concepts not encoded by single lexemes in the target language. Each component of the system "solves one part of the translation problem.

SAM calls the parser to translate each sentence into Conceptual Dependency. The interlingual representation is not syntactic; the syntactic "adjustment" rules, usually supplied in postediting, are part of the Mandarin surface generation routine. Hence, our system does not require editing of the input or output. SAM enables us to produce long or short paraphrases, summaries, or direct translations.

The Chinese generator is a modification of Goldman's BABEL.

Scripts influence lexical choices. The system can explain concepts for which no exact Chinese equivalent exists, using a network of conceptualizations to find related elements with single lexemes in Chinese.

For conceptual objects, a "generalized lexeme" can be constructed by following memory links to a token with a Chinese lexeme. The final component of the theory is a phrasal lexicon containing syntactic frames for modifiers and descriptors.

SESSION III

PLENARY LECTURE

TO BE ANNOUNCED

BERNARD VAUQUOIS

DYNAMIC PROCESSING IN QUESTION ANSWERING

WENDY LEHNERT

SAM answers questions about stories in restricted domains of knowledge; one issue which arises in SAM, but is relevant to all question answering, concerns two types of memory retrieval. Static response locates the information it needs in the memory representation that was generated when the text was read. Dynamic response gets information by actively reasoning from general world knowledge and inferencing in conjunction with the original representation.

SAM incorporates one dynamic response technique. 'Why' questions demand that a causality be identified; but we can ask for the causation behind acts that did not take place. Since the static memory representation does not embody information about nonactivities, some dynamic processing is needed. Such a question makes sense only if there is a possibility that the questioned event could have happened. SAM can generate paths of this type; to see these ghost paths, we point to the places where interferences or unusual things occurred and tell SAM to generate default paths from the immediately prior points. Once an act in question is found in a ghost path, we can answer by tracing the ghost path back up to the branch point and returning the interference or unusual occurrence.

Use of ghost paths is a dynamic response technique. Ghost paths do not carry information about everything that didn't happen, but I would claim that they contain everything you need to know. An intelligent program should know when a question is out of line; a program which has access to ghost paths will be able to recognize and process reasonable questions.

THE APPLICATION OF SCRIPT-BASED WORLD KNOWLEDGE IN AN INTEGRATED STORY-UNDERSTANDING SYSTEM

R. E. CULLINGFORD

This paper describes how traditional problems of reference specification, effects of context on understanding, selection of appropriate lexical items for generation, etc., are approached in a story understanding system using situational scripts. Emphasis in the design of SAM (Script Applier Mechanism) is on evolution of general principles which help to model, hence simulate, the processes which appear to go on when humans understand simple, script-based stories.

The version of SAM on which this paper is based consists of an analyzer to convert surface text into a Conceptual Dependency format; a script applier to build up a story representation, postprocessing programs to construct summary or paraphrase or answer questions; and a generator to render these structures into natural language. A co-routine control regime was selected to simulate close coordination among deep condeptual and input/output processes. We emphasize here the operation of the script applier.

SAM can handle references to three sources of knowledge in a uniform manner. "General knowledge about a situation" is contained in the script itself as a network of patterns with embedded roles. "Specific knowledge about a situation" becomes accessible when a situation is invoked. "Quasi-logical" knowledge is obtained by inferences.

The primary scriptal constructs used to control access to world knowledge are static context, a high-priority search list of patterns to match new inputs, and the script paths themselves. The paper discusses application of these constructs in interaction among modules.

APPLICATION DE TECHNIQUES RELEVANT DE L. INTELLIGENCE ARTIFICIELLE AU CODAGE ET A L'EXPLOITATION D'UN FICHIER DE RENSEIGNEMENTS BIOGRAPHIQUES MEDIEVAUX

MONIQUE ORNATO AND GIAN PIERO ZARRI

Le rassemblement systématique et la mise en forme de toutes informations biographiques sur les personnes impliquées dans les débuts du mouvement humaniste en France est l'une des taches de l'Equipe de Recherche sur l'Humanisme Francais des XIVe et XVe siècles. Sur le matériel recueilli, un projet de prise en charge globale sous forme de "mémoire sémantique" et d'exploitation par des techniques d'intelligence artificielle a été élaboré et soumis pour financement à la DGRST qui l'a accepté. Le choix de ce type d'outils est pour ainsi dire imposé par le caractère extrèmement complexe et très souvent implicite des relations interpersonnelles. De plus, le but visé est la création d'un système qui n'ait pas seulement une fonction statique de recupération d'une information stockée mais aussi une fonction dynamique permettant d'établir des liaisons nouvelles entre les données et d'accroitre en quelque sorte les connaissance de départ Cet aspect pratique se double de l'intéret methodologique d'évaluer sur un exemple concret le poids des outils à mettre en place pour simuler un ensemble de démarches intellectuelles d'une certaine complexité. La mémoire sémantique est organisée pour contenir deux niveaux de données: des information "personnages" et le métalangage qui sert de support à l'expression des informations. Les informations nouvelles qui vont alimenter le système sont regroupées selon des critères d'unité de temps et de thème et codées sous forme de plans dans le metalangage. Elles peuvent etre formulées dans un meme plan en fonction de plusieurs vedettes. Il a été prévue la mise au point d'un mode d'interrogation sur terminal qui admet la formulation de questions d'utilisateur et questions de système

A RESEARCHER FILE DESCRIPTION LANGUAGE AND ITS IMPLICATION IN INFORMATION RETRIEVAL SYSTEMS

SETSUO ARIKAWA

The researcher file is a collection of memos prepared through everyday research activities such as reading, discussion, attending lectures, and so on. At present the file is described at most in a tree structure, serving as a kind of thesaurus for document retrieval based on key words in logical formulas.

An ideal description language is a natural language such as English, French, and so on. Although studies by many scientists are proceeding, some essential difficulties are left unsolved; for a while we abandon natural language and propose a well-managed formal language RFDL defined on English.

Titles differ from sentences in English in that verbs are transformed into nouns or neglected as are subjects, but some prepositions are supplied. Stop words are useful for determining syntax and those derived from verbs for meaning. Fillmore s case grammar is useful.

The field to be dealt with is information science: information and control theory, automata, languages, and pattern recognition. Some functions for describing logical formulas in the first order predicate calculus and for concept formation are added to RFPL.

The formal systems are subsystems of that of Smullyan, stronger than CF and weaker than CS grammars.

In a new system researchers will be able to retrieve not only documents but also facts such as theorems, concepts, and data. As a byproduct the study of RFDL will also give a criterion for the writing of document abstracts.

SYNTAX AND FORMAL SEMANTICS OF ENGLISH IN PHLIQA1

S. P. J. LANDSBERGEN

In the PHLIQA1 question answering system several intermediate processing stages are distinguished. The formal languages of these stages are English-oriented Formal Language, World Model Language, and Data Base Language. This paper describes EFL and the transformation rules from English to EFL. Some theoretical aspects of EFL as a deep structure language are discussed.

EFL expressions are "trees"; a syntactic construction is a node from which labeled branches depart to subexpressions. The most important constructions are quantification, modification, function-application and nominal group. For every syntactic construction a rule specifies semantic types of immediate constituents and derivation of the expression's semantic type.

EFL is a formal language, not the somewhat hybrid, primarily syntactic, tree of generative semantics. A severe distinction is made between formal and referential semantics. EFL contains more constructions than predicate logic, and differs from it in the formal-referential distinction. Referential relations are functions; primitive predicates are formal.

SEMANTIC TYPES IN PHLIQA1

REMKO J. H. SCHA

The semantic representation languages are formal languages; the three differ mainly in their constants representing, in order, the terms and grammatical relations of English; the concepts of the Universe of Discourse; and the files and attributes of the data base and the available logical and arithmetic procedures. We want wellformedness in the syntax to include semantic meaningfulness; this restriction is achieved by means of the semantic types.

The type system is not just a classification of elementary objects; it contains many constructions for making "higher level" types. Most of these constructions can be nested arbitrarily.

We define functions which check whether types are equal, whether one type is included in another. A fragment of formal syntax is presented; conditions on and rules for computing semantic types are explained and motivated.

The computation of semantic types is useful in parsing and also in resolving polysemy at the level of the World Model. Semantic types are also used in the applicability conditions for semantic transformations which eliminate specific kinds of constants by reformulating the expression.

NIVEAUX D'INTERPRETATION DANS UNE TRADUCTION MULTILINGUE APPLICATION A L'ANALYSE DU RUSSE

N. NEDOBEJKINE

Toute analyse suppose la recherche de l'interprétation de la structure d'une phrase a un certain niveau. Ce niveau à atteindre dépend du degré de parente des structures de la langue-source et de la langue-cible. Plus celles-ci sont proche entre elles, moins le niveau recherché est haut. Nous avons retenu trois niveaux d'interprétation suivants:

Le niveau bas se limite à la combinaison de classes à l'intérieur d'une phrase.

Le niveau moyen suppose l'attribution à chaque groupe dans une phrase d'une fonction syntaxique traditionnelle, telle que sujet, objet, circonstanciel, détermination, etc.

Le haut niveau recherche des relations logiques assez précises entre les groupes (tels que agent, instrument, ... ou bien les cas profonds de Fillmore) et pourra convenir à la traduction entre langues suffisamment éloignées l'une de l'autre, ou bien au traitement des phrases à structures particulières dans les langues parentes.

La représentation de tous ces trois niveaux se fait au moyen de la meme structure arborescente dont les sommets contiennent des étiquettes relatives à chacun des niveaux d'interpretation ainsi que celles concernant les variables d'actualisation ou lexicales pouvant servir à n'importe quel niveau. Une règle syntaxique travaille sur ces arborescences en fonction des condition de schémas et d'informations grammaticales; elle transforme les premier et modifie les secondes.

Notre grammaire regroupe ses règles en trois sous-groupes rattachment de mots-oùtils, construction du groupe nominal, reconnaissance du groupe nominal entier et des constituants de la proposition autour du verbe.

JEUDEMO-CDC TO JEUDEMO-IBM

FRANCINE QUELLETTE

Since 1972, JEUDEMO, a package for producing indexes, concordances, and elementary statistics, has been in use at the Universite de Montreal. In collaboration with the research group of CNUCE, Pisa, we are implementing the first version on IBM equipment. In this paper we describe the conversion experience, which had five steps: (1) We made a preliminary study to find the main problems. (2) We held a working session at Pisa to explain the program to CNUCE's programmer who will work on the IBM version. During this four-week stage, we made the main conversion. (3) We executed our program in Montreal using the parameters of an IBM machine (word length, EBCDIC codes, etc.) producing extensive printouts at the main points of the program. (4) This version was sent to CNUCE to be reproduced and tested. (5) The CNUCE programmers made the adaptations necessary to optimize the program for an IBM machine. We are now planning another working session to determine what modifications of the program would make the software more useful to both centres, to decide what will be the steps necessary for the implementation of the final version of JEUDEMO on both computers, and to establish procedures for distribution of the software.

SESSION IV

PLENARY LECTURE

PROBLEMS OF INFERENCING AND ITS RELATION TO DECOMPOSITION

ARAVIND K. JOSHI AND STANLEY J. ROSENSCHEIN

In this paper. we will explore from various points of view some problems of inferencing, including the relation of inferencing to representation, and the relative ease of carrying out various operations in different representations. In particular, we will focus on the relationship of inferencing to decomposition of predicates into primitives, as well as other defined predicates.

This paper will be divided basically into two parts. Part I will consist of an analysis of the problem of inferencing along different dimensions. This will enable us to pull together certain key issues and their mutual relationships in a specific fashion. In Part II we will be concerned with the definitional hierarchies and the utility of setting up relations between defined predicates other than those that are implicit in their expansion into primitives. The design and implementation of a system of inferencing, whose major features are certain operations defined on a partial order over a set of patterns (schemas) has been described in print. We will suitably augment the set of these operations in order to provide a framework for our investigations.

PART I. Inferencing can be viewed along different dimensions, not necessarily completely orthogonal. Some of these dimensions are as follows

1. Direction of inferencing: Top-down (goal directed), bottom-up (data directed), or a combination of both. To what extent inferencing is "free running" or constrained by goals and subgoals.

2. Certainty of inferences: Certain, conditional, or conjectural.

3. Whether inferencing works with total or partial infor mation: Related to summarizing, which while accounting for all of the input forces one to imply more than what is in the input. Also related to lexicalization.

4. Criteria for controlling inferencing: Whether they are external or internal (i.e., structural). Related to the organiaation of the pattern space and how schemas are related to each other, e.g., in terms of shared information.

5. Domain dependent or independent.

6. Context dependent or independent.

7. Are the given facts structured and are the derived facts (inferences) integrated into the structure?

8. Does inferencing use definitional hierarchies or are all operations defined in terms of primitives into which all predicates are ultimately defined?

9. Is inferencing monotonic or nonmonotonic, i.e., does the addition of new schemas (with given inputs) always give at least the same inferences as we did before the new schemas were added (monotonic), or does it sometimes give fewer inferences than before (nonmonotonic)?

All these issues will be investigated in some detail; however, we will concentrate more on items 2, 7, 8, 9, and in particular item 8 will be investigated in much greater detail in Part II.

PART II. In defining predicates we have two choices: Either each predicate is defined directly in terms of primitives or it is defined in terms of other defined predicates. An example of the first approach can be found, but most studies in the decomposition of lexical items (verbs, in particular) follow the latter approach. Definitions in terms of other defined predicates may be set up for convenience (economy of representation) and ease of understanding; however, the question we want to investigate is the utility (with respect to certain operations) of postulating relations between defined predicates other than those implicit in the primitive expressions to which they can all be ultimately reduced. What are the trade offs between the two representations?

The analogy with programming languages is suggestive; programs are typically structured in terms of explicit hierarchies of subroutines, function calls, and so on. This provides advantages beyond intelligibility to programmers, namely saving of space, ease of reference, etc. The analogy is not exact, however, and the notion of hierarchies should be evaluated in light of the uses to which they are put

A difficulty with expansion to primitives is that sometimes appropriate inferences have to be made on the basis of certain combinations of primitives, i.e., in terms of some defined predicates. Responses of the system have to be often comparable to the input, e.g. responding to a question by an answer which contains unnecessary details is felt inappropriate

There is a disadvantage when we consider the problem of contradiction checking in the environment of a set of defined predicates, where each definition is a boolean combination of previously defined predicates. Unless the definitional system was constructed with great care, it will ordinarily be simpler to test for the existence of contradictory·

expressions in the expansion of an input set (given initially over the defined predicates) by expanding all of the inputs to primitives and doing the check at that level.

In the inverse process to expansion, namely SYNTHESIS of summarizing expressions, what are the advantages and disadvantages of explicit hierarchial structuring? WE FEEL THAT THERE IS A CLEAR ADVANTAGE IN EFFICIENCY OF THE SEARCH FOR FURTHER SUMMARIES, PARTICULARLY WHEN THE DEFINITIONS CONTAIN 'LOCAL VARIABLES' THAT CAN BE FILTERED OUT WHEN PROCEEDING TO HIGHER LEVELS OF THE HIERARCHY. This becomes particularly important when perceptual data (visual) is to be included in the input to be summarized. Even in synthesis, however, there is a drawback which should be understood. Alternate but equivalent definitions (equivalent in the sense of reducing to the same primitives) might cause relevant summaries to be missed. This issue has been raised by Bobrow and Norman, among others. Their suggestion for dealing with the problem is "consistent style" of definition. We will investigate how this can be insured automatically, if one allows auxiliary definitions.

COMMENT PERMETTRE AUX AVEUGLES D'ACCEDER AUX MOYENS MODERNES DE TRAITEMENT DE L INFORMATION

A. TRETIAKOFF

Je travaille en ce moment sur les problemes de communication aveugles-voyants et en particulier sur l'automatisation de l'utilisation de l ecriture Braille.

Durant ces travaux, des appareils permettant le stockage et le traitement de l'information en Braille ont été réalisés Ces appareils offrent de nouvelles possibilités pour la lecture, l'ecriture et le calcul en Braille. Ils permettent également aux aveugles d'accéder aux réseaux de transmission et traitement d'information en cours d'implantation dans de nombreux pays.

Je me propose de parler de ces diverses possibilités, et de leurs conséquences sur l'évolution du language Braille (abrégés, Braille international, etc.)

A COMPARISON OF TERM VALUE MEASUREMENTS FOR AUTOMATIC INDEXING

G. SALTON

A number of automatic indexing theories have been proposed over the last few years leading to the assignment of significance values to linguistic entities in accordance with their importance for purposes of content representation Among these are methodologies based on decision theory, information theory, communication theory, vector space transformation, and others.

An attempt is made to compare these theories by exhibiting the formal frequency characteristics which underlie them. The effectiveness of the various approaches is also evaluated in experimental situations by using collections of documents in the areas of aerodynamics, medicine, and world affairs.

SUPPORTING A COMPUTER-DIRECTED NATURAL LANGUAGE DIALOGUE FOR AUTOMATIC BUSINESS PROGRAMMING

GEORGE E. HEIDORN.

The user of a questionnaire-driven customizer need only answer a series of multiple-choice questions in order to obtain a business application program which is a version of a general program with parameters adjusted to his application. There is no interaction between system and user, if he does not understand a question, he has to look in an accompanying manual for relevant information

We are working on a dialogue customizer. The user's questions to the system may be about the application area, about the specific program being produced, or about the system itself Answers range from output of a prestored fact to partial simulation of the program being generated.

We have been observing actual customizer users completing questionnaires and also users of a manually simulated program explanation system, which we are automating by parts. The linguistic processing is done with NLP, using augmented PSGs. Decoding is bottom-up parallel; encoding is top-down sequential

A communications view of language is taken, rather than that of parsing and interpreting isolated sentences. User and computer engage in a dialogue with a certain amount of knowledge in common and help each other to know more. The system maintains a vector of context information and sets up expectations which in many cases simplify the analysis of user utterances. Much effort is being expended on the reference problem.

This paper gives both an overview of the project and specific details about the reference problem and dialogue context. The work is similar in many points to speech understanding.

INTERACTIVE ANALYSIS: A SYNERGISTIC APPROACH

DARYL K. GIBB

A powerful interlingua is the first requirement of a system that will accept a substantial number of English structures using a totally unrestricted universe (20 to 30,000 entries in lexicon). Such an interlingua should be able to show in an explicit way the differences among direction, selection, location, quantity, specificity, degree, number, contrast sets, manner, and so on. Analysis of syntax is often simplified to the point of recognizing word categories and perhaps case dependencies. The Junction Grammar interlingua is much more powerful and therefore logical deductions are often necessary to resolve ambiguities; the syntax of the interlingua is directly related to meaning--a subset or part of it.

The programs are designed to recognize potential ambiguous sentences and query a linguist as to the reading in the given context. Word sense and syntacto-semantic relationships can require resolution.

The interactive analysis produces a very explicit interlingua that can be manipulated and changed automatically if necessary before reconstruction in natural language. This type of system has several advantages: The linguist need not know any foreign language, since he answers questions about English. The system can generate input for programs which can be automatic, i.e. transfer and synthesis. It makes possible keeping records of interactions. Sometime in the future these records may be used as a guide in writing programs for automatic logic processing.

MODELING OF INDIVIDUAL SEMANTIC STRUCTURES

JAMES D. HOLLAN

In order to experimentally evaluate a given model of semantic memory, it is necessary to instantiate the model with some particular content. At this point an additional complexity is introduced. Performance in any task designed to test the psychological validity of a particular model might obtain from three categories of variables. structural, process, and content variables. Thus if one is interested in the effects of any one of these variables, it is necessary to control for the possible effects of the others.

The focus of research has been on structural and process variables with virtually no attempt to investigate or to control content variables. Allusions to the importance of content variables appear in the literature; in particular, to the possibility of performance differences between individuals due to differing semantic memory content, but only in one study have individual knowledge structures been constructed.

A technique has been developed to construct models of individual knowledge structures in accordance with a number of current memory models. The technique is implemented in the form of a PL/I program which effects the construction of the models and generates a graph-theoretical description of each model.

COMPUTER ACQUISITION OF NATURAL LANGUAGE: EXPERIMENTAL TESTS OF A PROPOSED SYSTEM

JANET KING, IAN MCMASTER, AND JEFFREY R. SAMPSON

This report sketches some highlights of the acquisition process as understood by linguists. It discusses methodological issues such as acoustic or orthographic input, grammatical formalism, nonlinguistic input, external environment, and cognitive development. It reviews computer-oriented natural language systems with acquisition components.

A new Complete Language Acquisition Program is proposed. CLAP's major components are a Perceiver, Semantic Base, Action Taker, Short-term Memory, Lexicon, Parser, Responder and components to modify parsing and responding strategies. It acquires language with five strategies sequentially: segmentation and meaning association, linear ordering, structural generalization, conflict resolution, and using discourse. CLAP emphasizes the primacy of comprehension over production and the role of a realistic external environment. At least the first three strategies are sufficiently well defined for immediate implementation.

Results are now available from two experimental implementations of part of CLAP's first strategy. The first learned the meanings of many object names. The second introduced actions and the verbs describing them. The first was influenced by Winograd, the second by Schank. Results of the second raised questions about methodology, including lack of concept-to-word linkages and the assumption that structural morphemes would develop no meaningful concept connections. Nevertheless, the second system learned many lexical items.

Further research will focus initially on implementation of the segmentation aspect of CLAP's first strategy.

SESSION V

PLENARY LECTURE

QUESTION AND ANSWER IN LINGUISTICS AND IN MAN-MACHINE COMMUNICATION

EVA HAJICOVA

Several treatments of question-answer relation are discussed (Belnap, Katz, Keenan, Conrad, and others), some oriented more linguistically, others more logically; requirements are sought which must be met by a theory underlying an effective question-answering system in man-machine communication.

It is argued that such a theory should take into account not only such conditions as presupposition sharing (where several levels should be distinguished, including corrective answers) but also the topic-focus characteristics of the question and the corresponding answers. The notion of focus of a question has already been applied in connection with man-machine communication by Winograd, but his views of the term focus differ from those of Halliday, Chomsky, and others

The topic and focus of questions are examined on the basis of the Prague School approach and it is shown under what conditions and to what extent the topic-focus structure of the question determines the form of the possible corresponding answers.

An experiment with a question-answering system is described, which is being prepared by the Prague group with the aim of building an automatic micro-encyclopedia in the field of electronics. The input consists (a) in several segments of English and Czech technical texts (chosen from monographs, papers, and entries from a technical encyclopedia); (b) in Czech and English questions concerning the relations between concepts of the given field; the input text is processed by a program of morphemic, syntactic and semantic analysis (i.e. translated into a "cognitive" language) and further by the brain of the system so that the concepts characterized in the input texts are properly stored in the data base and for every input question either an adequate answer is chosen, or it is stated that the information in question is missing in the system (and, if asked. for more frequently, it should be supplied into the system).

The output, yielded by programs of synthesis of English and Czech, translates the chosen answers from the cognitive language into the sentences of one of the two natural languages (according to the choice of the user).

ON INTENSIONAL AND EXTENSIONAL REASONING IN QUESTION-ANSWERING SYSTEMS

RAYMOND REITER

For concreteness, I shall focus on a query language very like that used by Woods, although any such first-order language would do. The end result of syntactic and semantic processing of a query is an expression in this language. Executing this expression extensionally answers the query. Such an expression often represents a highly inefficient call to the retrieval component. There may be nested quantifiers, multiple "such-that" conditions, paraphrase problems, and multiple data base representations of the same facts. I propose to specify an intensional description of how the data base is organized and use the description to select the best expression for execution.

Some queries are inherently nonextensional. An approach to queries in this class is to obtain a coarse description of an intention with respect to subsidiary functions and modify dynamically the value returned. A number of complicating problems can arise. In the paper I point out several such difficulties and propose techniques for dealing with some of them.

With hypothetical questions, it is by no means always obvious what functions are required. In such cases, some form of intensional reasoning on the hypothesis is necessary in order to identify it with an appropriate function call in the query language.

EXPERIMENTS IN CONCEPTUAL ANALYSIS OF THEORETICAL DISCOURSES

JEAN-GUY MEUNIER

By conceptual analysis, we mean the investigation of the semantical properties of a lexical form in a text. A concept is defined, in the fregean manner, as a function whose arguments will be a set of lexical properties. Being realized on a theoretical text, these conceptual analyses encounter original problems. In classical contemporary semantics, each word of a discourse receives one or more definite representations which which to understand the meaning of the text. This procedure follows the postulate that understanding must be related to knowledge, but it ignores the original contribution of text to meaning. In literary criticism, philosophical commentary, one cannot presume the meaning of the important words; the book has been written to define them.

We work on the French version of Descartes's Discours de la Methode. We try to discern the actual meanings of important words, semantical relations among them, and the distance of the author from the accepted meanings of the words.

Our strategy is creation of concordances, fragmentation of contexts, hypothetical definitions of words, and semantic preference analysis in the manner of Wilks.

When disambiguation and semantic selection cannot be operated, semantic formulas are transformed; this process continues until each keyword studied receives a satisfying definition in all its contexts. The contexts are then analyzed in a componential manner, and the keywords are compared for synonymy, conceptual inclusion, etc.

We hope that this research is an original application in a growing field of literary research by computer.

A NEW MORPH LEXICON FOR ENGLISH

M. S. HUNNICUTT

The lexicon, intended to facilitate the conversion of unrestricted text into speech, is comprehensive, useful in a variety of applications, and based on linguistic principles. The system includes a phonological rule algorithm and a terminal analog speech synthesizer. Future additions will allow for the production of natural-sounding speech at the sentence level. These additions, now existing as separate modules, are algorithms which generate a surface structure parse and govern fundamental frequency, duration, and timing.

Motivating factors include the desire to model the process used by a native speaker while reading; the comprehensiveness of a morph lexicon, and efficient use of memory.

The lexicon was obtained by decomposing 50,406 distinct words found in a corpus of 10^6 running words. Beginning with a base of 1, 2, and 3-letter words and a decomposition algorithm the lexicon was built up by adding to the base all n-letter words which did not decompose into words of less than n letters. The algorithm uses a recursive longest-match-first procedure from the right end of the word and has a set of morphophonemic rules. for suffixing, including plurals and palatalizations among others.

Since the first decomposition found by the algorithm was not necessarily the correct one, a set of selectional rules was devised.

Polymorphemic words remain in the lexicon as required for the conversion of text to speech. These entries are annotated. The lexicon is of potential interest to lexicographers, to linguists and to anyone in need of a large data base of English words.

CAN SOME PROCESSES OF LANGUAGE EVOLUTION BE SIMULATED?

BERND S. MÜLLER

The algorithmic approach undertaken in the "triangle world" (T) tests some very crude hypotheses. The evolutionary process has the following conditions: Beings who are supposed to develop a language exist in a world which consists of their habitations and a food-producing outside world. Food has triangular square, and other regular 2-dimensional forms; Triangles are the most tasty food. A set of "world rules" tell about the edibility of forms other than triangles and the possibilities for the tranformation of nontriangles into triangles. Eventually different tribes in T describe their outside world differently when their possibilities to transform geometric forms are restricted.

In general, language evolution takes place in a world which is governed by certain world rules; parts of the world are the language-developing beings which experience situations and communicate in steadily changing types of languages; the languages are build up from random signs for specific world phenomena; the changes from one type of language to another are caused by certain language evolutionary rules which seem to be mostly economic in nature; sign structure rules seem to belong to the language evolutionary rules.

In T, the language evolutionary rules are identity, abstraction, differentiation, and preference for short signs.

The evolution process produces language according to meta-rules and world rules. The evolution product becomes more refined if the set of world rules does. Simulation of the T type can only produce a language with some of the most general features of a human language.

This paper describes the construction of the T program, its productions and results. It contains preliminary reflections on more refined types of evolution models including those of the stochastico-algorithmic type.

ON ALGEBRAIC DISTRIBUTIONAL ANALYSIS OF ROMANIAN LEXICAL UNITS

LIANA SCHWARTZ POPA-BURCA

The equivalence relation generated by Dobrusin's domination relation gives rise to corresponding equivalence classes which coincide with distributional classes. We have studied the contextual behavior of the written Romanian verbal, nominal, and adjectival forms. It has been proposed to perform alge braic distributional analyses by employing several levels of grammaticalness, such that each of them contain the previous ones. A level of grammaticalness is just a finite set of contextual classes; it is introduced to emphasize some contextual peculiarities of the elements in a chosèn corpus. The existence of elements having more than one grammatical valence causes most of the problems related to contextual equivalence in Romanian.

We discuss here some aspects of algebraic distributional analysis of Romanian verbal forms, comparing them with nominal or adjectival forms. Five levels are considered. We obtianed 76 distributional classes.

Every grammatical category at the first level is a collection of simple forms of the indicative corresponding to one and only one grammatical person. At the second level, the subjunctive appears. At the third, the past and present participle, and some infinitives.

The method does not restrict choice or order of levels of grammaticalness; such a problem is a false one. If we fix the first level and perform distributional analysis by taking into account more than two levels as well as the same aspects of contextual behavior, the order of subsequent levels does not affect the result of our analysis.

ALGEBRAIC DISTRIBUTIONAL ANALYSIS OF CERTAIN FRENCH WORDS

LUCREȚIA VASILESCU

The main aims of our research are to establish distribution classes and the relation of domination between them and the elementary grammatical categories and types of homonymy. The analyzed words are the noun, the adjective, and the verb.

We obtain the correspondence, from the point of view of semantics, for the whole of the nouns, adjectives and respectively the French verbs. We also obtained a regularization and a new distribution of the parts of the sentence. The so-called exceptions were given their own law.

A measure of the morphological homonymy appearing in the paradigms of grammatical categories is given by the index of morphological homonymy. According to this index, nouns and adjectives are organized by couples of two or three noun or adjective forms, this behavior being found within the distribution classes and the elementary grammatical categories as well. With respect to the verb, the number of elementary categories is much bigger, mainly beacuse of the graphical aspect varying from one person to the other.

The behavior of these words has not been considered exhaustively, the analysis can be refined by introducing other classes of contexts, as well as ordering and choice criteria for grammatical levels.

SYSTEME INFORMATIQUE POUR LA GENERATION MORPHOLOGIQUE DE LANGUES NATURELLES EN ETATS FINIS

Benoit Thouin

L'objet de cette communication est un système informatique -package- pour la generation morphologique de langues naturelles destiné à compléter la chaine de traduction automatique du G.E.T.A. Le systeme recoit en entree une arborescence représentant la structure syntaxique d une phrase ou d'un ensemble de phrases. Il donne en sortie la chaine finale correspondante. Formellement, le systeme est un transducteur arborescence-chaine compose de deux transducteurs. Un automate d'èxecution simule les transducteurs; l utilisateur a la responsabilité de lui fournir les informations propre a chaque unité lexicale et les regles de generation morphologique qui seront appliquées. Cette communication de donnees au systeme se fait en quatre temps au moyen d'un langage specialise: Déclaration des variables; declaration des formats et conditions; écriture des dictionnaires; Ecriture de la grammaire de génération morphologique.

Les etats de l'automate d'execution sont les noms des règles; l'etat initial est le nom de la première regle dont la condition est satisfaite par le masque en entree. L'automate s'arrete pour un masque donné dès qu'il n'y a plus de règle suivante, ou si une règle suivante exigee n'est pas applicable. Le système est suffisamment flexible pour laisser au linguiste le choix de sa strategie de génération. De plus il est assez puissant pour exprimer des phenomenes particuliers tels que l'élision, la contraction et la formation de mot composes. Enfin le systeme est conversationnel--implanté sous CP/67-CMS--ce qui permet de continuels retours dans la définition des variables, dictionnaires et regles appliquées.

DETECTION AUTOMATIQUE DES VARIATIONS ORTHOGRAPHIQUES SUR DES NOM PROPRES--DEFINITION D'UN TRANSDUCTEUR MORPHO-PHONETIQUE INTERACTIF

YVES CHIARAMELLA

Nous présentons ici un outil permettant de definir de manière interactive tout modèle de transduction des mots sous forme de chaines de symbole phonétique (transduction morpho-phonétique), ainsi qu'une application relative à l interprétation phonétique de noms propres extraits de documents anciens.

Les variations des caractéristiques phonologiques ainsi que celles des règles particulières sont particulierement importantes dans les applications de Démographie historique. La plupart des individus concernés n'ayant qu'une connaissance orale de leur nom, les orthographes correspondantes présentent de nombreuse variations. On peut definir trois niveaux de rapprochement des variations orthographique sur des criteres d'ordre purement phonétique, ceci nous permet de définir sur l'ensemble des noms une partition hiérarchisée et de là, une mesure de ressemblance entre les noms.

Notre outil de base est l'analyseur morphologique du système PIAF; qui a pu etre adapté au role de transducteur phonétique tout en conservant ses propriétés fondamentales Le programme PIAFPHO en est un dérivé orienté vers la classification automatique de mots sur des critères de proximité phonétique. Dans le cas qui nous concerne, nous avons défini trois niveaux de classification hiérarchisée correspondant à autant de modèles de transduction phonétique. Enfin, une grande souplesse a été prevue au niveau du mode d'entrée des données et de sortie des résultats. entrées et sorties peuvent etre effectuées indépendamment sur console ou sur fichiers magnétiques, ce dernier mode permettant le traitement en masse des données (plus de 5 000 noms dans notre application)

SESSION VI

PLENARY LECTURE

ON CONTEXT-FREE PARSING

B.A. SHEIL

The importance of context free languages for the description of natural language phenomena has long been recognized, and automata which accept the context free languages form an integral part of many natural language systems. However, the non context free aspects of language require that such automata be not directly applied to natural language, but that their underlying principles be abstracted and incorporated into the designs of more general processors. Thus, it is surprising that so little work on abstractions of the context free parsing problem has been done by computational linguists This paper reports the results of such an investigation which are strikingly at variance with widely held beliefs on the subject.

The major evaluation criteria for any algorithm are the amounts of time and space it requires for its worst case. Thus, the first question is what aspects of a context free parser allow it to achieve polynomial, rather than exponential, parsing (the limiting case achieved by enumerating all finite derivations). Although many different properties (including "parallel" searching, the avoidance of backtracking, etc.) have been proposed, it is shown that one such property, use of a well formed substring table both holds for all known polynomial parsers, and can be shown to be sufficient in and of itself to produce polynomial behavior. (A parser has the wellformed substring property iff the results of analyzing any substring of the input in terms of some nonterminal of the grammar are recorded so that such an analysis is performed at most once, irrespective of how many times the analysis may be used during the parsing.) The proof proceeds by showing that the search space for such a parser is polynomially bounded, without reference to the order in which it is searched. Furthermore, the specific bound placed on the search space allows the cubic bound for Chomsky Normal Form grammars, and the quadratic bound for linear grammars to be shown as corollaries of this result.

The WFS result is very surprising given the wide range of algorithms that have been proposed to achieve this effect. It specifically refutes the widespread conjecture that backtracking parsers are inherently exponential. It implies that other aspects of the algorithm may be chosen independently to optimize other aspects of performance, while the WFS preserves the polynomial bound. One is tempted by its invariable presence in parsers that achieve this performance to conjecture that it is necessary as well as sufficient, but such conjectures are very difficult to establish.

Given the polynomial bound, and in particular the cubic bound for any given cfl, the next major issue is the ability of the algorithm to achieve tighter bounds for restricted classes of the cfgs. Thus, it is often asserted that a major advantage of the Earley algorithm is that its bounds for unambiguous and LR(k) grammars are quadratic and linear, respectively. However, it is argued that these are really two quite separate issues. As it can be decided by inspection whether a grammar is LR(k) for any given k, it is clear that any syntactic system which desires linear performance on this class is able to achieve it by special casing. Nor can the cost of this special casing be held against this strategy as the same amount of inspection is required by Earley's algorithm to determine the correct lookahead parameter. On the other hand, as there is no procedure for determining whether an arbitrary cfg is ambiguous, it is highly desirable that the same parser used in the general case have quadratic behavior on an unambiguous grammar.

On this issue, unlike the previous one, the bounds cannot be established independently of the sequence in which the algorithm traverses the search space. Although it can be shown that the successful parse must lie in a quadratically bounded space, an algorithm that searches top down (i.e. one that considers a constituent before establishing the satisfiability of its subconstituents) does not necessarily confine its attention to this space. However, algorithms which form constituents in a bottom up fashion, i.e. those that consider a constituent only after establishing the satisfiability of its subconstituents, can easily be shown to be so confined..

The third issue discussed is the preprocessing of the grammar required by the parsing algorithm. Once again, although it has been claimed that avoidance of such preprocessing is a major advantage of some algorithms (e.g. Earley's), no basis can be found for this. While it is possible to construct algorithms which depend on extreme deformation of the productions of the grammar (e.g. into Greibach Normal Form) which make it difficult to reconstruct the constituent structure of the original grammar, it simply does not follow that all deformations of the grammar produce such problems. In particular, parses represented in a Chomsky Normal Form (the main target of this critique in Earley's paper) can be converted to the constituent structure of the original grammar in real time, making the representation of the grammar used internally by the parser completely transparent.

The ideas presented are illustrated by the construction of a new context free parser--the recursive descent parser-which is a simple top down, depth first, backtracking algorithm which uses a WFS table both to achieve cubic bounded parsing and to prevent cycles on left recursive productions An extremely simple proof of correctness and confirmation of the bounds predicted by the general theorem are presented. Simple extensions to the algorithm allow the derivation of quadratic bounds for unambiguous grammars, and linear bounds for a class that includes the finite state and palindrome grammars. (The top down strategy precludes linear bounds for the LR(k) grammars but, as outlined above, this is not considered a drawback.) Furthermore, the use of depth first search results in very good performance on highly ambiguous grammars. Consequently, although the worst case is still cubic, the parser rarely approaches this bound for inputs accepted by the grammar.

Three major conclusions are drawn from this study. First is the paramount importance of the WFS table for any algorithm dealing with languages with context free subsets. The strong evidence for its necessity and sufficiency for polynomial parsing indicates that natural language systems should strive for a structure that permits efficient use of this device. Second is the irrelevance of many of the issues that have been claimed to be of major importance in this area. Finally, based on these conclusions, the recursive descent algorithm was developed. Because of its top down approach and its close parallel to the generative model of context free grammar, it is easily both understood and proven correct. Because of its use of the WFS table and depth first search, it is as efficient as any parsing algorithm known. For both these reasons it is suggested both as a pedagogical tool and as a practical context free parser.

THE TRAVEL BUDGET MANAGER'S ASSISTANT

BERTRAM BRUCE AND B. L. NASH-WEBBER

The program is a vehicle for studying natural language text and speech understanding. Its task is to aid in planning and allocating money for trips. It needs a diverse array of knowledge about acoustic signals, phonetics, syntax, semantics, travel budgets, etc. For this system to be comprehensible, debuggable, and capable of improvement, it must be clean, understandable and efficient in organization.

Each component can be developed and tested independently, yet can interact conveniently with the others. The flow of information between components is explicit.

Four components (Syntax, Semantic Interpreter, Retrieval, and Audio-Response Generator), together with the System Controller, function as a complete text understanding system. In processing spoken input, the System Controller activates the real-time interface to acquire the signal, then the signal processing component to compute parameters, then the Acoustic Phonetic Recognition component to produce a segment lattice for input to the Lexical Retrieval component. Control then passes to the Speech Understanding Controller which uses the Syntactic, Semantic, Lexical Retrieval, and Verification components to arrive at a model of the utterance.

In this paper we consider explicit representation of interaction and information flow among components; isolation of factors which determine how and when interactions should occur; and evaluation of component effectiveness.

A MULTIPROCESSING MODEL FOR NATURAL LANGUAGE PROCESSING

R. SMITH AND F. RAWSON

The system evaluates the meaning of natural-language sentences of informal mathematics. In this domain some representational problems are less severe than elsewhere, and we assume that evaluation of the meaning of an utterance is the determination of the logical form in a manner suitable for a proof checking system applied to computer-assisted instruction.

The problem we focus on is how to handle scopes of quantifiers and operators in paraphrases of mathematical formulas. We propose to evaluate sentences by associating a separate process with each node of the surface-level syntax tree. We believe that it is sufficient to have a context-free grammar for the surface syntax. We give a detailed outline of the proposed implementation of a LISP-like language, PLISP, that we are designing. The language includes primitives for accessing and creating processes; it combines features of SAIL and the TENEX timesharing system.

The paper shows how PLISP functions can be written to handle some natural language paraphrases of mathematical formulas, including function application, pronouns, and quantifiers. The method is not however limited to this domain.

Few of our insights about scopes and operators are particularly new. Our objection to the methods of e.g. Woods and Winograd is that the information about the role that an operator plays becomes too globally distributed in the code of the program and is hence difficult to describe an a way that clarifies the understanding of natural language processing. Transfor mational grammar can be interpreted as directed toward the solution of the problem that concerns us, but determination of "deep structure" requires more than syntactic information. Also, some inverse transformations create evaluation difficulties that are resolved by PLISP.

A COMPUTERIZED SUPPLEMENT TO THE DICTIONARY OF MIDDLE DUTCH

F. DE TOLLENAERE

It is no surprise that the Middle Dutch Dictionary of E. Verwijs and J. Verdam (1882-1929) should call for revision. In 1965, Dr. J. J. Mak accepted a commission to compose a supplement; he retired in 1973; the 19 card boxes of material collected were transferred to the Leiden Institute for Dutch Lexicology.

The material is being punched on paper tape for transfer to disk. It will be alphabetized and printed out on continuous form. Several small files are being processed; some large ones will be treated later.

Once the list is complete, it could be transferred to magnetic tape to be printed by photocomposition. The printed list could then be edited as a separate little volume, or added to the concise Middle Dutch Dictionary of Verdam

Although our Institute will not edit the Supplement Verdam and Mak once hoped to produce, the supplement material collected by both Verdam and Mak will at least become accessible. It is beyond doubt that it will only constitute a modest makeshift, but one which may be useful for the study of Middle Dutch.

A METHOD FOR A NORMALIZATION AND A POSSIBLE ALGORITHMIC TREATMENT OF DEFINITIONS IN THE ITALIAN DICTIONARY

N. CALZOLARI AND L. MORETTI

Our aim is to define, in line with an intensionally oriented semantic theory, a formal representation of the noun definition set taken from the Dizionario della Lingua Italiana (Zingarelli, Bologna, 1970), which has about 120,000 entries.

The method is inductive--to reach an enucleation of 'semantic markers' and the 'relations' between them only on the basis of the dictionary definitions of lexical items. The dictionary definitions show a certain trivial regularity; it is easy to isolate a generic and a specific part. The high-frequency words in definitions, other than syntactic words, are mostly nouns; moreover, the nouns most often quoted as semantic markers in the literature on the subject

A network structure is proposed for representation. The generic part of a definition should correspond to a path on an oriented graph, the nodes labeled with 'markers'; the relations will be few, mostly functions or relations' These 'relations' will be the algorithms of the graph itself. In the specific part of a definition, a pointer to a lexical entry is allowed.

PROBLEMES ACTUELS EN TA: UN ESSAI DE REPONSE

CH. BOITET

Les systemes réalisès jusqu'ici utilisent une sémantique rudimentaire et une organisation figée en une succession de phases prédéfinies. Améliorer de tels systèmes, c'est introduire l'utilisation de méthodes heuristiques et adaptatives, permettre une interaction entre les différents niveaux, et se servir d'une sémantique plus élabòrée.

L'organisation du système GETA est sequentielle: un fragment de texte est traite successivement par ses quatre composants, puis le systeme passe au fragment suivant.

En TAUPHA, l'analyseur morphologique du GETA est modifie pour permettre la construction d'un graphe de chaines parallèlement à l'analyse, la correction des formes non reconnues, et l'appel sur un certain nombre de formes, ou jusqu a un marquant.

On définit un nouveau composant ALGOG. Pour lui, chaque sommet de l'arbre des choix est une analyse partielle et certains sommets terminaux sont des analyses complètes. Un second composant nouveau MONIT permet de realiser une interaction entre les composants, et d'échapper a la stricte organisation du traitement en phases successives. Ceci est possible parce que c'est au niveau de MONIT qu'on définit l'algorithme d'analyse syntaxique (comme une heuristique) L'ALGOG est capable d'une adaptation qui consiste à munir les arcs de poids qui évoluent au cours du traitement (apprentissage "paramétrique")

Il est porbable qu'une sémantique "référentielle", permettant des inférences par simples règles de transformations de réseaux, ait une valeur. Il faudrait savoir aussi manipuler efficacement des structure récursives. Cette étude reste à faire.

DESIGN AND IMPLEMENTATION OF AN ENGLISH-FRENCH TRANSFER GRAMMAR

RICHARD KITTREDGE, LAURENT BOURBEAU, AND PIERRE ISABEL

The TAUM group has designed a distinct transfer grammar which expresses the correspondences between nuclear sentences of two langauges and between the transformations which can be applied to construct more complex sentences in each language. Here a transformation is a mapping between surface structures which preserves acceptability ordering among the sentences which have that surface structure. There is often a one-to-one correspondence between the transformations of English and French. Nuclear sentences also show a greater similarity between languages than do complex sentences

The overall procedure for translation is as follows. Nuclear sentences are normally translated by finding translations for the nonderived nouns: Then the predicate words are translated as a function of the noun subcategories. We work up the parse tree, calculating the French transformation(s) which should correspond to each English transformation or combination of them. The transfer is complete when the topmost transformation of the English structure has been used to calculate a corresponding French transformation.

In the current TAUM 76 system, transfer rules are separated into distinct modules for each transformation class where the correspondence between languages is not 1 1. Comprehensive modules are being tested for passive article, and tense transfer, and up the list of possible lexical translations of an English word for the syntactic class which the analysis assigns to that word. The proper translation is the first in the list which satisfies the conditions on the structural context. In some cases the final choice is delayed until the full target structure has been calculated.

TOWARD A QUANTITATIVE HISTORY OF ENGLISH POETRY: PRELIMINARY RESULTS

COLIN MARTINDALE

This paper describes computerized content analytic studies of 88 English poets born between 1490 and 1950, undertaken to test a theory of literary history: The role of poet includes a force leading toward change, the necessity to produce original works. The pressure for novelty leads to changes in style and content that can be predicted psychologically; to be more original, one must regress. Regression is limited; at some point, stylistic rules must change.

Dictionaries of regressive and concrete imagery and of semantic differential scores are used, with statistical analysis programs

Analyses reveal a number of statistically significant results. The indices of primary process content and of concreteness and imagery exhibit a clearly sinusoidal upward-moving trend, as predicted. There is weaker evidence for increases in incongruity and lexical diversity. Data for stylistic change per se are not yet available. These preliminary results are seen as being supportive of the theory. Plans for further analysis of the corpus and for collection of series of non-literary texts for control purposes are described.

SESSION VII

PLENARY LECTURE

TITLE TO BE ANNOUNCED

MARTIN KAY

ORGANIZATION AND CONTROL OF SYNTACTIC, SEMANTIC, INFERENTIAL AND WORLD KNOWLEDGE FOR LANGUAGE UNDERSTANDING

F. Hayes-Roth and D. J. Mostow

This paper describes a taxonomy of knowledge types and a related scheme for knowledge organization and computational control: a uniform framework in which to embed the diverse sorts of knowledge and behavior which are apparently essential for complex language understanding tasks. Our basic assumptions are (1) Each unit of knowledge may contribute information; (2) Each unit is probabilistically errorful and it is unknown a priori whether use of knowledge will generate helpful results; (3) The number of potential contributions vastly exceeds the requisite minimum for understanding. Our method is to identify general types of knowledge-based behaviors, to construct systems that can recognize data patterns where such behaviors are justified, and to control order of computation so that behaviors which appear most helpful are computed first. The four types of behavior rules are recognition, hypothesization ("prediction"), enumeration ("respelling") and postdiction.

The four have the same data-driven form (precondition, response) The function of knowledge-based inferencing is to generate and support hypotheses. These observations suggest a clean and simple structure for language understanding systems. Systems organized as proposed should exhibit increased uniformity, controllability, extensibility, and transferability.

COMPUTATIONAL EXPLICATION OF INTENSIONALITY

JANUSZ STANISLAW BIEN

An expression is intensional if it can be transformed into a nonequivalent expression by replacing one of its members by an expression which is equivalent to that member. In the present paper I discuss only one type of intensional expression; i.e. reported speech, but the treatment can be generalized to other cases of intensionality.

One trivial and obviously not adequate solution to the problem of substitution is to disallow substitutions in the indirect context, e.g. by treating the reported sentence as a name. It is natural to define the extensional equivalence of programming language expressions as the equaiity of the values delivered by evaluation of given expressions; from this definition it follows that programming languages are intensional. All the variables which can be accessed by an expression together with the values of these variables are called the environment of the expression.

We treat natural language utterances to be run in our brains. The environment contains the data bases representing the knowledge and abilities of the person. In every moment of discourse we have at least one environment, that of the narrator; in reported speech we have a choice of at least two, since that of the person referred to is available. The pivot' of the reported sentence is evaluated in the second environment, but in almost all cases the definite descriptions can be evaluated in either.

A multiple environment framework solves the problem of reported speech in a strict and intuitive way. The approach will be applied by the present writer to other cases of intensionality.

SYNONYMIE LEXICALE: UNE TENTATIVE D'ANALYSE

AMEDEO CAPPELLI

Ce travail utilise les définitions des entrées du Dictionnaire de Machine de l'Italien, soit pour etudier, d'un point de vue théorique, la synonymie, soit, plus particulièrement, pour systématiser ces définitions dans le but d'une organisation plus générale du DM.

On a établi une procédure de génération d'arbres de synonymes Les résultats de notre analyse montrent que les unités que nous avons classifiées comme synonymes ont, entre elles, des rapports qui ne sont pas seulement de synonymie mais aussi de hyponomie, hyperonymie, etc. Cela nous permet de préciser l'inadéquation de la théorie lexicale qui a été à la base de l'élaboration des données que nous avons utilisées et nous force, évidemmènt, à les systématiser.

Nous faisons cette systématisation sur la base d'une théorie qui décrit le lexique en termes de rapports de synonymie, etc. Le moyen par lequel nous vérifions cès rapports est essentiellement un test de nature syntaxique; en d'autres termes, les unités lexicale sont analysées en les insérant à l'intérieur de particulières phrases.

Pour établir l acceptabilité de ces phrases nous avons pris en considération surtout une analyse basée sur l'intuition. D'autre moyens de vérification consistent dans la comparaison des unités sur la base de leurs définitions analytiques. On utilise aussi d'autres codes, relatifs aux définitions, qui sont déjà contenus dans le DM. Il s'agit de codes relatifs aux usages, comme, par exemple, archaique, rare, figuré, etc.

MECHANICAL RESOLUTION OF LEXICAL AMBIGUITY IN COHERENT TEXT: ALGORITHMS AND EXPERIMENTAL RESULTS

Y. CHOUEKA AND F. DREIZIN

Experiment shows that a Hebrew word can have on the average about 4-5 different interpretations, not counting slight semantical variants, figurative meanings, etc. The causes are lack of vowels in writing and addition of prepositional, conjunctional, and pronominal elements to a wordform, hence the possibility of different decompositions. The second cause is examined here.

Our working hypothesis is that if in a coherent text several words can be analyzed as having the same meaning (stem, or generally, dictionary entry), then they should be so analyzed.

Algorithms were constructed and applied to a few documents of the rabbinic medieval "Responsa" literature; three results were immediately apparent. (1) For word forms of consistent equivalence classes the algorithms were almost 100% correct. (2) Most equivalence classes are consistent. Excluding the "common words" and a few cliches, then with very few exceptions the equivalence classes that remain are consistent. (3) With respect to the set of roots rather than stems that can be realized in a given word form, the number of exceptions is practically reduced to zero.

ANALYSIS OF JAPANESE SENTENCES BY USING SEMANTIC AND CONTEXTUAL INFORMATION

M. NAGAO AND J. TSUJII

The parser of our question-answering system transforms fairly complex Japanese sentences into abstract structures marked for case; it uses detailed semantic descriptions in the dictionary and contextual information extracted from the preceding sentences. For the present, we confine it to the domain of elementary chemistry where we can describe the semantic world in rather concrete terms but where complex events occur: disappearance, emergence, and change of properties do not seem to occur in Winograd's block world.

We classify nouns into categories of 'entity', 'attribute', 'value', 'action','prepositional' and 'anaphoric' From these categories, 16 semantically acceptable pairs occur.

When we find a conjunctive postposition, we search out the word in the following string with deepest semantic similarity to the head preceding the postposition.

The meanings of analyzed sentences are represented in the form of a semantic network (Simmons et al.). We search it to find words for empty case elements. A trap list holds pending questions until later sentences can answer them

The parser is an ATN; semantic and contextual functions are programmed in LISP 1.6. Results for sentences from a junior high school chemistry textbook range around 90%.

CONSTRUCTION D'UN DISPOSITIF EXPERIMENTAL POUR LA REPRESENTATION ET LE TRAITEMENT DES DONNEES TEXTUELL ES ILLUSTRE SUR UN ÉXEMPLE EN HISTOIRE

EUGENE CHOURAQUI AND JACQUES VIRBEL

Cette communication s'appuie sur une recherche plus globale visant à définir les éléments d'une méthode d'investigation pour l'analyse des données textuelles utilisant les méthodes et les moyens de l'information. Un examen approfondi de la conduite d'analyse d'objets textuels, tant du point de vue des méthodes relevant des sciences humaines--linguistique, histoire, etc.--que de celles relevant de l'informatique, nous a conduit à situer notre démarche méthodologique par rapport à celle des sciences d'observation.

Dan les termes d'un resume bibliographique habituel, l'expérience particuliere qui est prise comme illustration a consiste a vérifier si un matériel textuel donné, les inscriptions funéraires des vétérans de l'armée romaine trouvées en Afrique du Nord, pouvait etre daté selon des méthodes de sériation s'appuyant sur une description du contenu de ces textes.

Le dispositif expérimental visant à vérifier cette hypothèse a été décomposé en un certain nombre d'étàts expérimentaux: (A) Formulation du problème historique posé et définition d'une méthode de résolution à partir d'un ensemble d'hypotheses. (B) Définition des collections d'objets textuels correspondant à cette démonstration. (C) Définition des corpus de représentations abstraites et formelles des textes. (D) Constitution du domaine de definition du traitement des objets formels. (E) Resolution formelle du problème pose. (F) Interpretation et validation des résultats formels. (G) Répercussions sur le dispositif expérimental et conséquences dans le domaine historique.

A MODEL FOR FUNDAMENTAL FREQUENCY CONTOURS IN ENGLISH

JONATHAN ALLEN

In this paper, we integrate together a wide range of factors which determine English fundamental frequency (Fo) contours so as to permit the algorithmic determination of these contours from a linguistic description of the utterance. We start by regarding every sentence as having a (possibly deleted) performative verb which characterizes the speaker's intent and the illocutionary force of the utterance. The performative contains within it an S node, which dominates the sentential nucleus and operators.

The nucleus contains the basic ideational proposition of the sentence and is characterized in the Fo contour by a slowly falling curve, modulated by accents on the semantically important content words and segmental effects due to vowel tongue height and the voiced-unvoiced nature of consonants before syllabics. Indeed, it is just these effect which have been noted as the Fo correlates of simple declarative sentential utterances in English.

The content of the proposition, however, is but one of the communicative functions of the speech act marked in the Fo contour. We present extensive evidence that modality items, used to represent the speaker's attachment to the truth value of the proposition, are characteristically marked by Fo since they fulfill the interpersonal function in the communicative act. An extensive corpus, including both sentence and paragraph materials, was read by three speakers. This corpus provided systematic variation of modal auxiliaries, negatives, subject quantifiers, and sentential adverbs. These sentences, such as "Some of the boys might not have studied their books", show marked Fo protrusions on the modality items, indicating that the speaker uses this means to inform the listener(s) of his attitude toward the truth value of the basic underlying proposition. Since these effects are largely independent of the position of the modality item, they can be predicted from knowledge of the presence alone of these interpersonal markers

Once the ideational and interpersonal communicative functions have been represented in the Fo contour, it still remains to include textual or discourse effects due to focus-shifting transformations and shared knowledge with the listener(s) To study these phenomena, another corpus was recorded by three speakers. The focus-shifting transformations included passive, there-insertion, clefting, pseudo-clefting, topicalization, right and left dislocation, extraposition, adverb preposing, and gapping. Examples include "A carrot was eaten by the farmer." and "Never has the farmer eaten carrots." Once again, we show that each of these transformations is associated with a characteristic Fo gesture, which is utilized by the speaker to display to the listener(s) the focus of the utterance. Within this corpus, the effect of new and old information is also studied, including repeated items, pronouns, and ellipsis. As expected, it is the new items that form the focus and receive Fo accent.

From these studies, a comprehensive model for Fo contours is derived, which accounts for a wide variety of speech act phenomena, as described above. In order to derive natural sounding speech using synthesis by rule, it is felt that all of these factors must be systematically included, and that together they form a cohesive linguistically motivated model for Fo contours in English.

HIERARCHY OF SIMILARITIES BETWEEN PHRASE STRUCTURES

PIERLUIGI DELLA VIGNA AND CARLO GHEZZI

It is not decidable whether two grammars are weakly equivalent, and it is decidable whether two grammars are strongly equivalent; but even strong equivalence is too dependent on the grammars (an extra renaming rule in one grammar falsifie equivalence) and fails to capture the notion of similarity between different languages to be paired even in very simple translations.

We propose a four-level hierarchy of similarity: (1) structural equivalence; (2) identity of languages, "similarity" of structure (subtrees of fixed maximum length instead of single productions correspond to the syntax trees); (3) no constraints on the sentences generated; only structural similarity is involved; (4) a permutation can occur in the correspondence between the subtrees which leave two corresponding nodes of the syntax trees-

All four levels are decidable. A decision algorithm is given which is almost the same for the four levels, except that the constraints to be taken into account at some steps are of different strength. Moreover, the algorithm gives, as a secondary result, the possibility of rewriting the two given grammars, if similar, in such a way that the structural equivalence implies a 1:1 correspondence between the rules of the rewritten grammars.

INDEX OF CONTRIBUTORS

INDEX OF CONTRIBUTORS

BRUCE, Bertram 64
Bolt Beranek and Newman Inc
50 Moulton Street
Cambridge, Massachusetts 02138

BURTON, Richard R. 16
Bolt Beranek and Newman Inc.
50 Moulton Street
Cambridge, Massachusetts 02138

CALZOLARI, N. 67
Divisione Linguistica del CNUCE
Pisa, Italy

CAPPELLI, Amedeo 74

CHAUCHE, J. 25
Groupe d'Etudes pour la Traduction Automatique
Boite Postale 53
F-38 041 Grenoble Cedex, France

CHIARAMELLA, Yves 57
Laboratoire d'Informatique
U.S.M.G.
Boite Postale No 53
38041 Grenoble Cedex, France

CHOUEKA, v 75
Institute of Information Retrieval and Computational Linguistics
Bar-Ilan University
Ramat-Gan, Israel

CHOURAQUI, Eugene 77
Laboratoire d'Informatique pour les Sciences de l'Homme
Centre National de la Recherche Scientifique
France

COURTIN, Jacques 18
Laboratoire d'Informatique
U.S.M.G.
Boite Postale No 53
38041 Grenoble Cedex, France

CULLINGFORD, R. E. 31
Yale University
New Haven, Connecticut

DELLA VIGNA, Pierluigi 80
Istituto di Elettrotecnica ed Elettronica
Politecnico di Milano
Piazza L. Da Vinci 32
Milano, Italy

DE TOLLENAERE, F 66
Institute for Dutch Lexicology
Leiden, Netherlands

DREIZIN, Felix 75
Institute of Information Retrieval and Computational Linguistics
Bar-Ilan University
Ramat-Gan, Israel

FERRARI, G 24

FILLMORE, Charles 10
Department of Linguistics
University of California
Berkeley
GERASIMOV, V. N. 8
USSR

HUNNICUTT, M. S. 52
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ISABEL, Pierre 69
Traduction Automatique
Universite de Montreal, Canada

JOSHI, Aravind K. 38
Department of Computer and Information Science
Moore School of Electrical Engineering
University of Pennsylvania
Philadelphia 19174

KAY, Martin 9
Xerox Palo Alto Research Center

KING, Janet 47
Department of Computing Science
University of Alberta
Edmonton, Canada

KITTREDGE, Richard
Traduction Automatique
Universite de Montreal, Canada

LANDSBERGEN, S. P. J. 34
Philips Research Laboratories
Eindhoven, The Netherlands

LEHNERT, Wendy 30
Department of Computer Science
Yale University
New Haven, Connecticut

ROSENSCHEIN, Stanley J. 38
Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York 10012

ROSS, Donald Jr. 15
English Department
University of Minnesota
Minneapolis 55455

SALKOFF, Morris 23
Universite de Paris VII, LADL
2, Place Jussieu
Paris 5e, France

SALTON, G. 43
Department of Computer Science
Cornell University
Ithaca, New York 14853

SAMPSON, Jeffrey R. 47
Department of Computing Science
University of Alberta
Edmonton, Canada

SCHA`, Remko J. H. 35
Philips Research Laboratories
Eindhoven, The Netherlands

SGALL, Petr 11
Charles University
Prague, Czechoslovakia

SHEIL, B. A. 58
1250 William James Hall
Harvard University
Cambridge Massachusetts 02138

SMITH, R. 65
Institute for Mathematical Studies in the Social Sciences
Stanford University, California 94305

STUTZMAN, Walter J. 28
Artificial Intelligence Project
Yale University
New Haven, Connecticut

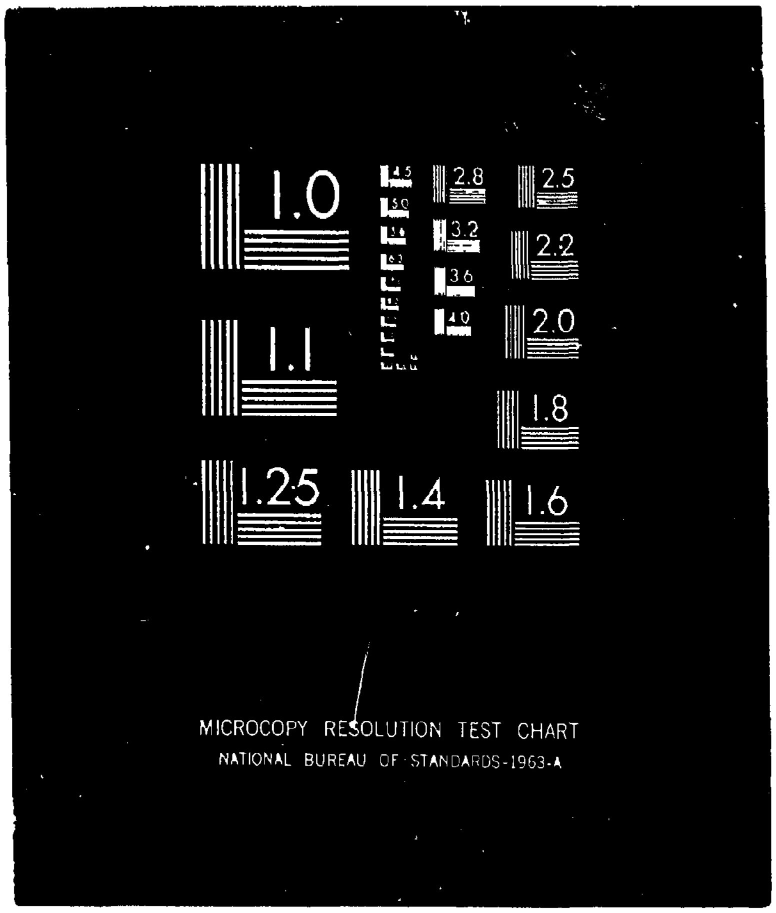
TAKEYA, Shun-Ichi 17
Research Institute of Fundamental Information Science
Kyushu University
Fukuoka 812, Japan

THOUIN, Benoit 56
Groupe d Etudes pour la Traduction Automatique
Universite Scientifique et Medicale de Grenoble, France

THOMPSON, Henry 20
Department of Linguistics
University of California
Berkeley

TRETIAKOFF A. 42
Electronique Linguistique Informatique Appliquees
43, avenue Lulli
92330 Sceaux, France

END


1.0
4.5
5.0
5.6
6.3
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A