American Journal of Computational Linguistics *Microfiche 50*

THE FINITE STRING

NEWSLETTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS

VOLUME 13 - NUMBER 5 MAY 1976

TABLE OF CONTENTS

An organization for a dictionary of word senses
Dick H. Fredericksen 2
Current Bibliography 24
Cahiers du groupe de travail Analyse et experimentation dans les sciences de l'homme par les methodes informathiques - E. Chouraqui and J. Virbel 93
Bibliography and subject index, current computing . . . 94
Directory of university computer science 95
Privacy, security, and the information processing industry - Dahl A. Gerberick 96

AMERICAN JOURNAL OF COMPUTATIONAL LINGUISTICS is published by the Center for Applied Linguistics for the Association for Computational Linguistics.

EDITOR: David G. Hays *Professor of Linguistics, SUNY Buffalo*

EDITORIAL ASSISTANT: William Benzon

EDITORIAL ADDRESS: *Twin Willows, Wanakah, New York 14075*

MANAGING EDITOR: A. HOOD ROBERTS *Deputy Director, Center for Applied Linguistics*

MANAGEMENT ASSISTANT: **James Megginson**

PRODUCTION AND SUBSCRIPTION ADDRESS: *1611 North Kent Street, Arlington, Virginia 22209*

American Journal of Computational Linguistics *Microfiche 50 : 2*

AN ORGANIZATION FOR A DICTIONARY OF WORD SENSES

DICK H. FREDERICKSEN
IBM THOMAS J. WATSON RESEARCH CENTER
YORKTWON HEIGHTS, NEW YORK 10598

ABSTRACT: This paper describes a lexical organization in which "senses" are represented in their own right, along with "words" and "phrases", by distinct data items. The objective of the scheme is to facilitate recognition and employment of synonyms and stock phrases by programs which process natural language. Besides presenting the proposed organization, the paper characterizes the lexical "senses" which result.

1. Introduction.

This paper describes an internal lexical organization which is particularly designed to capture the facts about synonymy. Besides recording the inclusion of each word in one or more synonym sets (identified with its various "senses"), the scheme attempts to distribute attributes perspicuously between "senses", "wordings", and the intersections of the two. In addition, there is provision to record multi-word idioms, stock phrases, and the like, and to include these as elements in synonym sets when appropriate.

Briefly, "senses" are represented in their own right, along with "words" and "phrases", by distinct data items. Each word or phrase is associated with a list of the "senses" which it can express; conversely, each "sense" is associated with a list of "alternative wo lings" Additionally, each word is associated with a list of phrases in which it occurs.

Grammatical category, features, selection restrictions, and the like are applicable at three different levels: to words or phrases as such, to "senses" as such, or to particular usages of words or phrases (equivalently, to particular wordings of "senses").

This lexical organization has been implemented at IBM Research, Yorktown Heights, N.Y., by a program -- not to be described here -- which builds such dictionaries in a very compact form, giving interactive assistance to the person making the entries. (For example, the program points out the possibility of merging "senses" whenever their wordings overlap and their attributes are compatible, and merges them if so directed.) There are suitable facilities for saving the results, retrieving them in various ways, and for altering such things as schemes of classification without scrapping previously prepared work.

The ultimate intent is that the "dictionary of senses" should serve as the lexical component in a natural language fact-retrieval system. Pending its incorporation in that role, it will be used to amass and organize information on the semantic relations among words and phrases.

The balance of this paper comes in two sections:

Section 2 presents the proposed lexical data structures, and suggests how they are to be used. Included is a sketch of how various types of grammatical and semantic "attributes" fit into the scheme.

Section 3 discusses the character of the "senses" encoded in the resulting dictionary. Reasons are advanced for regarding lexical "senses" as something far short of semantic primitives. At the same time, synonym sets are defended against the view that "true paraphrases are rare or nonexistent".

2. The Internal Representation.

It will be our purpose in this section to say just enough about internal representation to lay bare the organizing principles of the lexicon. The focus is on architecture and motivations; details of field layouts, internal codes, etc. are not at issue here.

To make the discussion concrete, suppose we are interested in the senses of the word "change" Assuming that none of the words are unfamiliar, the following should put us in mind of two senses:

change: 1. *v* alter;
2. *n* small coin.

This, of course, is just a dictionary entry in the traditional format (though with synonyms offered in lieu of definitions).

On the other hand, we might approach the same information from a different direction: starting with the two concepts, we might seek words to express them. It is difficult to picture this latter situation without assigning artificial labels to the concepts. Call them concepts 1 and 2, and suppose for a moment that there were a practical way to look the concepts up *(without* having thought of either word for either concept). Then the information to be retrieved might be envisioned this way:

1. *v* change, alter
2. *n* change, small coin

It is this duality of viewpoint -- that words have senses, while senses have wordings -- that our lexical representation must reflect.

The starting point, then, is that words, phrases, and "senses" are separately represented. There are three principal types of data item, plus a standard connector:

1. A *"Key Data Item"* (KDI) represents a single word.

2. A *"Phrase Data Item"* (PDI) represents a string of two or more words which are to serve as a unit in some context.

3. A *"Sense Data Item"* (SDI) represents one distinct sense common to a set of words and/or phrases. In general, a word or phrase may be usable in more than one sense, while a given sense may have alternative (synonymous) wordings. Both these types of variability are recorded making use of the next data item:

4. A *"Sense Link Element"* (SLE) is a connective item, to be explained shortly.

Three principal fields will engage our attention in each type of data item. Fig. 1 summarizes the fields for each type.

"Alternative Senses" Link	Global Attributes	"Phrase Involvements" Link

KDI (Key Data Item)

"Alternative Senses" Link	Global Attributes	"Component Word" Links

PDI (Phrase Data Item)

"Alternative Wordings" Link	Global Attributes	"Sense Chain" Link

SDI (Sense Data Item)

"Alternative Senses" Link	Local Attributes	"Alternative Wordings" Link

SLE (Sense Link Element)

Fig. 1
Schematic of Data Items, with Principal Contents

Each KDI (Key Data Item) or PDI (Phrase Data Item) contains an "alternative senses" link -- a pointer to the first SLE (Sense Link Element) in a chain of SLE's which represent the various senses of the word or phrase. The SLE's are chained via their own "alternative senses" links, and the final member points back to the KDI or PDI. Thus, we shall speak of such a chain as a ring specifically, an "alternative senses ring". If no senses are on record for a particular word or phra , the "alternative senses" link in the KDI or PDI is self referent.

Reciprocally, each SDI (Sense Data Item) contains an "alternative wordings" link. This leads to a chain of SLE's which represent more-or-less synonymous wordings that express the sense. These SLE's are chained through their own "alternative wording links, and again the chain is closed into a ring -- this time beginning and ending with the SDI.

The structure that is shaping up may now be seen in Fig. 2. *The crucial point is that each SLE represents the intersection between an "alternative senses ring and an "alternative wordings" ring.* From the standpoint of the word or phrase, it represents a particular sense; from the standpoint of the sense, it represents a particular wording

Starting from a KDI or PDI, one gets to the SDI for a particular sense by advancing along the "alternative senses" ring to the relevant SLE, then detouring along the ring which connects the latter to the SDI (as one of the SDI's "alternative wordings"). Starting from an SDI, one gets to a particular wording by the reverse process. Since each "alternative senses" ring contains exactly one KDI or PDI, while each "alternative wordings" ring contains exactly one SDI, each SLE is tied to exactly one sense of one word or phrase. (Equivalently, it is tied to one wording of one sense.)

The next point of interest is that "attribute" fields are present in all four types of data item -- even in the connectors (SLE's). The attributes which may be recorded in each, however, come from different bags.

To begin with, the attributes found in an SDI characterize all the wordings of a given sense whenever the wordings are used in that sense. In Fig. 2, for example, sense "1" should be marked as a "verb" sense, while sense "2" is a "noun". One would not wish to record the attribute "verb" in the KDI for the word "change", for the KDI represents facts about the word itself, irrespective of sense, and "verb" does not hold for all uses of the word "change". On the other hand, "verb" does characterize all wordings of sense "1", whenever they're being employed to express that sense. It would furthermore apply to any additional wordings which we might think of, such as "modify", provided they are really used in a synonymous way.

As a matter of fact, it turns out that the traditional parts of speech -- noun, verb, adjective, preposition, etc. -- fit best in this scheme as global attributes of senses, recorded in the SDI's.

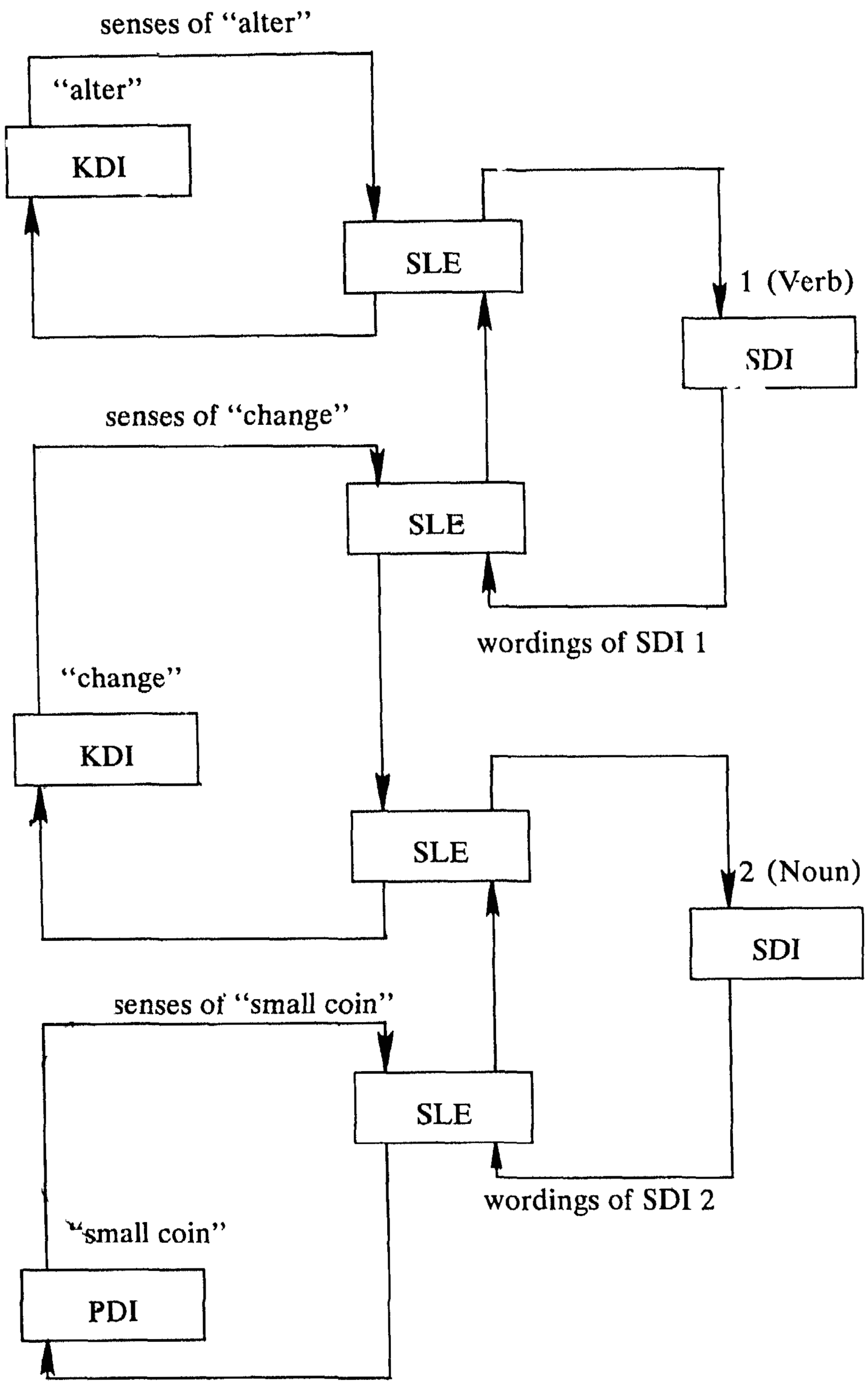


Fig. 2
"Alternative Senses" and "Alternative Wordings" Rings

(The first sense has two wordings: "alter" and "change". The second sense has wordings "change" and "small coin". Two senses are recorded for "change", and one sense each for "alter" and "small coin".)

A different sort of attribute may be recorded in a KDI, as a global feature of the word itself. For example, we may note of the word "change" that it is "regularly conjugated". That is, when used as a verb, it forms the third person singular by adding "s", and both past and past participle by adding "ed" To be sure, this "global" attribute applies only to the "verb" senses of "change"; but a moment's reflection will confirm that "change" has more than one "verb" sense, and the regularity of its conjugation is common to all of them. Thus, it is useful to note this regularity as an attribute of the word itself. (Contrast this with the behavior of the word "can", which is regular when it means "to pack in cans", but irregular when it means "is able to".)

Various other attributes suggest themselves as global characterizers of the words themselves, to be recorded in the KDI's. For example, one might wish to note of "change" that it drops its final "e" when adding "ing" (this is the normal rule) but of "singe" that it doesn't.

Still other attributes are appropriate when characterizing multi-word units (in PDI's). A string of words whose meaning is not evident from the mere juxtaposition of its constituents (such as "give up") may be classified as an "idiom". A string of words whose meaning could be figured out from the meanings of its constituents, but which occurs with enough frequency to warrant inclusion in the dictionary, might be classed as a "stock phrase". (Example: "drop dead".) A string like "perform in a subordinate role", which one would not normally expect to encounter in its own right, might be classed as a "definition" (for a certain sense of the word "accompany", difficult to reword except with a definition).

Perhaps the most unexpected site for recording attributes is in the connective elements (SLE's). These are the logical place, though, to note features that apply to a specific sense of a word, without being global to either the sense or the word. Consider the following four sentences:

> On the way to the office, he stopped daydreaming.
> On the way to the office, he ceased daydreaming.
> On the way to the office, he ceased to daydream.
>
> versus:
>
> On the way to the office, he stopped to daydream.

Suppose we choose to view this as a restriction upon the (surface) object of the verb: "stop", when applied to an action, must take a gerund as its object; "cease" can take either a gerund or an infinitive. (It wouldn't affect the point being made if we said that "stop" inhibits a certain grammatical transformation en route to surface structure, while "cease" permits it.)

Now, we wouldn't want to mark "gerund object only" as a global attribute of the sense, for we have just shown that "cease" and "stop", two wordings of the sense, differ with respect to this restriction. On the other hand, it doesn't belong among the global attributes of the word "stop" as such, for "stop" has other verb senses, even transitive ones, to which the restriction is completely inapplicable. (Consider "stop a hole in the dike", "stop a catastrophe", etc.) That leaves the alternative we are suggesting: treat the restriction as an attribute of one particular usage of the word (equivalently, one particular wording of the sense).

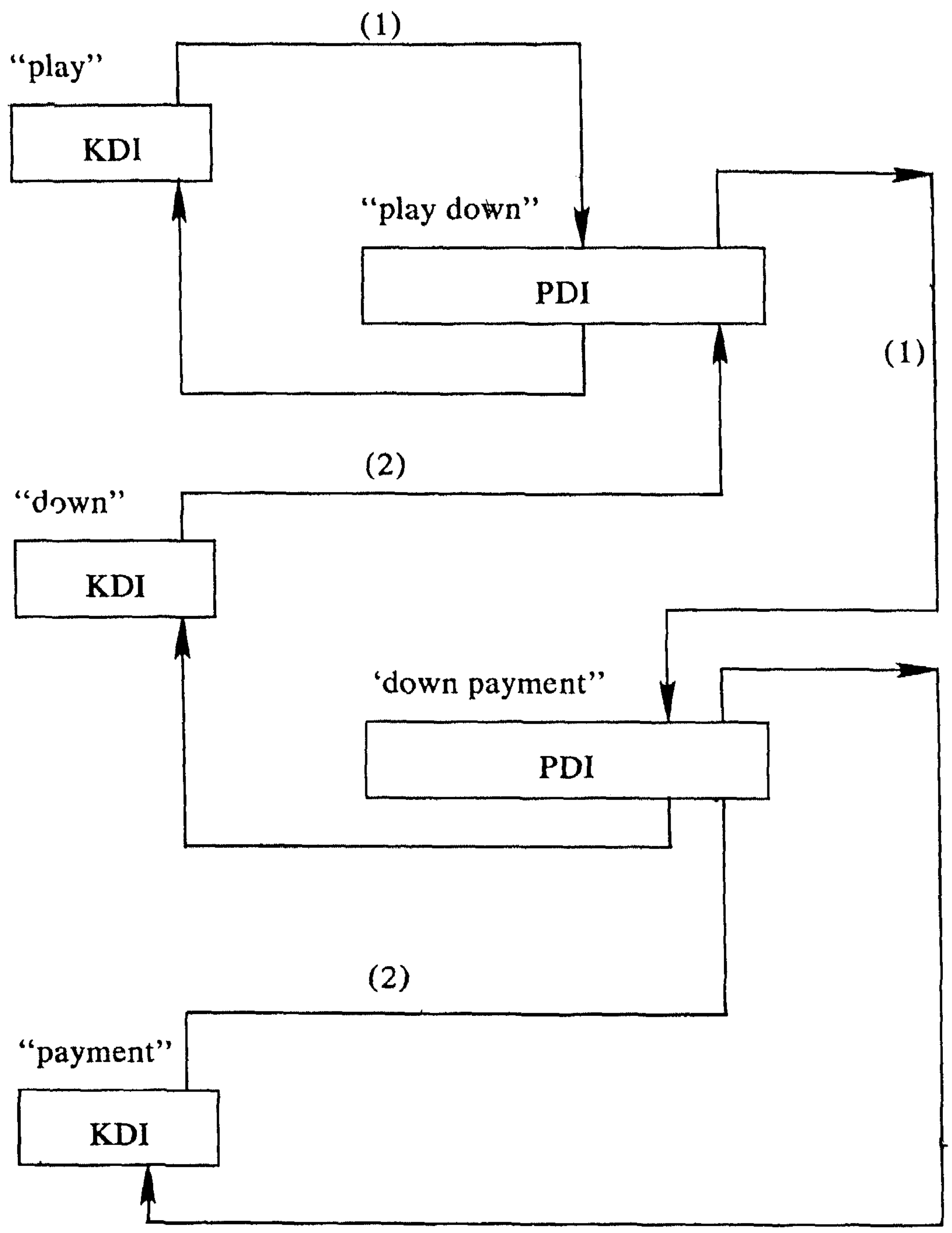


Fig. 3
"Phrase Involvement" Rings

(Where numbers are shown on connecting links, they indicate the position of the word in the phrase which is linked to.)

Besides having senses, individual words are involved in phrases, and this fact is also represented in our data structure. Fig. 3 shows the plan of attack. In the KDI for each word, there is a link connecting it to the PDI for the first phrase in which the word is known to occur, together with a number designating the position of the word (1st, 2nd, 3rd, etc.) in that phrase. In the PDI itself, there is a continuation link for each word of the phrase, together with its number in the next phrase. In the final PDI involving a given word, the link for that word points back to the KDI. Thus, independent of its "alternative senses" ring, each KDI may have a "phrase involvements" ring.

This structure makes it possible to retrieve all the idioms, stock phrases, definitions, etc., in which a given word has made its appearance, anywhere in the dictionary. As the same structure is used to encode every multi-word unit, no occurrence of a word is ever lost sight of, and a phrase can be looked up via any of its constituent words.

Of the fields to which Fig. 1 calls attention, we have discussed all but one. In the SDI for each "sense", there is a "sense chain" link field. This links the SDI to its successor in a global chain of "senses". Using this chain, it is possible to make an exhaustive, non-duplicative list of all the "senses" recorded in the dictionary. The listing program has only to proceed down the chain, retrieve from each SDI its attributes, decode them, then chase around the "alternative wordings' ring of the SDI and list the wordings alongside the attributes.

One more feature of the internal representation deserves mention: the data items for words occur as "leaves" in a lexical tree (Fig. 4). That is, the KDI for a word can be looked up letter-by-letter, following a chain of pointers that correspond to successive letters. The chain ends at a KDI after following a substring sufficient to distinguish the word from the nearest thing like it in the dictionary. The lexical tree has the advantage that words can be looked up either at random or in sequence.

Recapitulating, these are the essential features of the representation:

*1) "Senses" are represented separately from "wordings", and the mutual connections between them are made explicit in both directions.

*2) "Wordings" may be either single words or multi-word phrases. These are represented by distinct types of data item, and may be subject to distinct schemes of classification, but they are on the same footing with regard to "sense" connections. With each word is associated an exhaustive list of the phrases in which it occurs.

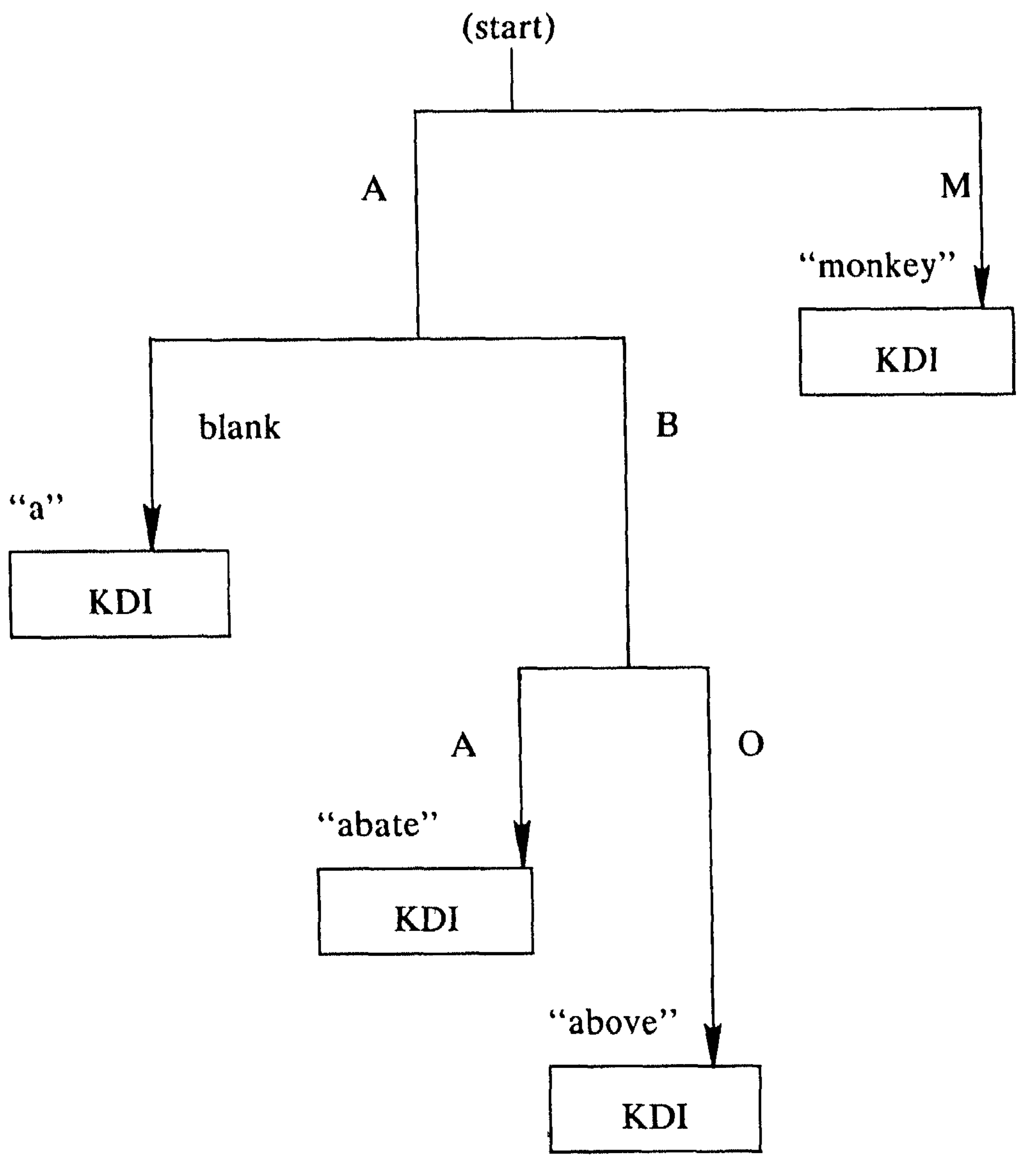


Fig. 4
Lexical Tree

(For a dictionary containing only the words "a", "above", "abate", and "monkey", this would be the full tree. The path to each word is only as long as needed to distinguish it from the neighbor with which it shares the longest leading substring.)

*3) Classifiers and features, drawn from appropriate sets, may be attributed separately to words, to phrases, to senses, or to particular senses of words or phrases (i.e., to particular wordings of senses).

*4) The data items which represent senses are globally chained, and may be exhaustively listed.

*5) The data items which represent words are accessible as "leaves" of a lexical tree; hence they may either be retrieved by lookup (in response to presentation of the words) or volunteered in alphabetical order.

Given a commitment to represent a lexicon as suggested by points *1 through *5 above, various implementations would be possible. Alternative implementations of individual points (though not of the scheme as a whole) have in fact been described by other writers. The lexical tree (*5), for example, is no great novelty: Sydney M Lamb and William H. Jacobsen describe implementation details of one such tree [5]. [10] also concerns a dictionary which uses this general style of organization for lookup. For that matter, the lexical tree is reminiscent of Feigenbaum's "discrimination tree." [1]

More interestingly, the separate representation of senses and wordings has been incorporated in other systems by R. F. Simmons ([11], [12]) and by Larry R. Harris [3]. This way of looking at matters led Harris to remark some of the same points that we have been stressing: that senses have alternative wordings just as words have alternative senses; that multi-word phrases might occur on the same footing as individual words in the expression of a sense; and (interestingly enough) that part-of-speech information really adheres to the "sense", not to the "word" Similarly, Simmons associates his "deep case" information with lexical nodes representing "wordsenses", while words themselves are treated as "print image" attributes of the wordsenses.

Harris's dictionary was only a minor component in a small-scale model of concept acquisition. No great number of either words or concepts was required to illustrate the principles at stake, so Harris programmed the dictionary as an array, with words represented by rows and "concepts" by columns. Elements of the array were merely frequencies, indicating the strength of association between each word and each concept.

Needless to say, for a full-scale vocabulary of words and concepts, such an array is mostly empty; nobody would dream of expanding it in that form. From a programming standpoint, the only thinkable choice is some form of list structure. Having decided in principle to use "some form of list structure", though, one might well ask: Why chains? Why rings? Why not just include in each Key Data Item a full list of pointers to the corresponding Sense Data Items, and vice-versa?

The answer is simply one of convenience. It's easier to handle insertions and deletions when they don't require the movement of expanded items to new quarters, or the provision of "overflow" pointers. It's easier to reclaim freed storage when deleted items come in a handful of standard sizes. As for "rings", they eliminate the need for two-way pointers, since one can break into a ring at any point and follow it to its source.

It should be noted that to make rings an attractive representation, the details of the material being represented must cooperate. In particular, the rings must not become too long, or the processing required to follow them becomes excessive. It happens that "alternative senses" rings and "alternative wordings" rings are typically short rarely more than a dozen links per ring. "Phrase involvement" rings, on the other hand, can become spectacularly long, especially for words like "a" and "to". In practice, it's necessary to provide these rings with short-cut links.

Any of these programming details could be altered, however, without abandoning the essence of the scheme, which is given in points *1 through *5 above.

3. The Character of Lexical Senses.

Perhaps the first thing to get straight about the "senses" represented in this dictionary is what they are *not*. They are not "concepts"; they are not a set of "primitives" into which human experience an be decomposed. No conjecture is put forward here that any such collection of discrete, atomic concepts even exists, let alone that it might be finite.

Rather, the "senses" of the dictionary are in the nature of fuzzy equivalence sets among words. (This is only a metaphor; we shall do more and more violence to the technical notion of an "equivalence set" as we proceed.) Each "sense" groups a set of words which, in a set of appropriate contexts, might be used more or less interchangeably. That the equivalence sets are fuzzy, one can convince oneself with but the briefest immersion in the materials of the language -- trying to decide whether particular words belong in particular groups or justify the creation of new groups.

Consider, for example, the following set of words and phrases:

(abandon, give up, surrender, relinquish, let go, desert, leave, forsake, abdicate)

Clearly, there is a common theme that can run through all of these, given the right circumstances. It might be expressed as "reluctant parting from somebody or something". This can be seen by coupling the verbs with various possible objects:

(abandon, give up, surrender) a town to the enemy

(abandon, give up) all hope

(give up, relinquish) one's claim to an estate

(give up, let go) our entire stock at a loss

(abandon, desert, leave) one's wife and children

(desert, forsake) a friend in need

(give up, abdicate) the throne

(abandon, desert) an exhausted mine

(forsake, give up) all other, keeping thee only to her/him

(abandon, desert, leave) the area threatened by the storm

Should we, then, declare this group of words to be a "sense"? There are difficulties. The various words carry nuances, which it may or may not be easy to ignore in a particular context. "Forsake", for example, can suggest that there is something reprehensible about the action. It can also connote formal renunciation, and the above example from a marriage vow shows that the formality can be present without the reprehensibility. Nuances get in the way of interchangeability; it would sound strange to substitute "desert" into the marriage vow.

Besides nuances, the individual words have conventional areas of application. One does not normally say that the doctors "deserted" all hope, or that an errant husband "surrendered" his wife and children. The minister officiating at a wedding would be considered daft if he adjured the bride and groom to "abdicate" all others, and a merchant would not advertize that he was "relinquishing" his entire stock at a loss. (Somehow, the latter situation calls for more pedestrian language.)

At the opposite extreme, overawed by this lack of interchangeability, we might decide to respect the unique personality of each word, abolishing equivalence classes altogether. The inconvenience of such a cop-out is obvious: we then have to introduce some *other* mechanism for recognizing the equivalence of utterances that are intended synonymously, though they employ different words. But beyond being inconvenient, the exclusion of equivalence sets is a denial of linguistic facts -- just as bad, in its own way, as the naive attribution of unconditional synonymy.

For it is a commonplace of everyone's experience that the speaker and the listener agree to ignore the nuances of words, whenever nuances get in the way of communication. A writer who has used the word "give up" eight times in five lines will surely cast about for some alternative ways of saying the same thing. If "relinquish" and "abandon" would normally be too flowery, or if "surrender" would in other circumstances call to mind an armistice ceremony in a railway wagon, that will not deter the writer from tossing in a few occurrences of those words -- once a context has been established that discourages the overtones. Nor will the reader understand matters any differently. It is as if writer and reader conspired: "We're fed up with *that* word, let's hear another." Or, perhaps, the writer simply connives at jolting the reader awake with frequent changes of idiom, maybe even an occasional incongruity. In any case, synonymy is imposed upon the words, and this literary behavior merely exaggerates what people do habitually in common speech.

Not only can words be stripped of nuances normally present; they can *take on* colorations suggested by the context. The suggestion of "reluctance" conveyed by all the verbs of our example can be inferred, in at least one case, from the setting alone; and in this case, a variety of more neutral verbs could be used synonymously:

(part with, take leave of) our entire stock at a loss

One could even substitute the word "sell", and it wouldn't change the meaning that was already bad into the utterance. But to admit context-dependent synonymy of this degree is to stretch the equivalence sets" to the point of uselessness.

It comes to this: neither the grouping nor the separation of words can be fully justified. Grouping is nearly always conditional, and separation is often so. If one could anticipate all possible contexts in which a group of words could occur, one could perhaps enumerate all possible equivalence sets -- one for each combination of word group with a set of contexts making the words interchangeable. Anyone, however, can see the futility of that aspiration.

In the end, one settles for messy compromises. Words are grouped if a largish set of contexts in which they are interchangeable springs readily to mind. They are separated (into perhaps overlapping groups) if the imagination readily suggests contexts in which their meanings differ "significantly" -- whatever "significantly" may mean. In doubtful cases, when words are grouped somewhat questionably, one promises oneself to add markings some day that will prevent misuse of the equivalence. When words are separated somewhat questionably, one promises oneself to add a mechanism some day that will recognize their relatedness.

In the end, too, one assigns internal structure to the equivalence sets. That's the effect of assigning local attributes to the alternative wordings ("animate subject", "object a vehicle", etc.): constraints are imposed upon the interchangeability of the wordings. More radical structuring can be accomplished if, for example, one notes "government" as an alternative wording of the sense "govern, rule, control", with the attribute "nominalization".

A trenchant discussion of such difficulties may be found in Kelly and Stone [4]. There the emphasis is upon disambiguation: given a word in a passage of text, they seek to identify (by selection from a fixed list of possibilities) the sense in which it is used. Building a computerized dictionary for the purpose, they soon became concerned with the arbitariness and the proliferation of target "senses", as taken from standard desk dictionaries. They argue, with persuasive examples, that what lexicographers conventionally distinguish as separate senses of a word are often just applications of the word's underlying concept to different contexts. To cover the various contexts, the underlying concept has to be stretched a little, by a process of metaphoric extension. This metaphoric process is beyond our present power to computerize, but for the long run looks indispensable for successful language processing. Meanwhile, the authors advocate a dictionary which records for each word as few discrete senses as practicable, combining into one sense all the usages which can reasonably be united by a common underlying thought.

It is interesting to re-examine Kelly and Stone's argument with a different task in mind: not the disambiguation of one word, but the recognition of synonymy between two words. A metaphorical capability would be as useful for the one task as for the other, but in the case of synonym recognition, some of the considerations which have guided traditional lexicography remain pertinent. In particular, it is necessary to ask not merely whether the concepts overlap, but whether the one word may in fact be used in place of the other. As noted before, usage is restricted by conventional domains of application; for example, an "alteration" is conceptually both a "change" and a "modification", but one wouldn't *call* it a change or a modification when painting a sign for a tailor's shop.

The arbitrariness of the equivalence sets is not all that disqualifies them as "conceptual primitives". There is a much deeper difficulty in the fact that practically all "senses" can be paraphrased in terms of other "senses". Take, for example, the intransitive sense of "change" (as in "My, but you've changed!"). Surely, one would suppose, the concept of "change" must be primitive? Change of state is what well-nigh a third of all verbs are about.

But if "change" is a "primitive", it's a peculiar sort of "primitive", for it can be paraphrased in a variety of ways:

> (change, become different, cease to be the same, assume new characteristics, make a transition into a new state)

Note that the multi-word paraphrasals are not idioms; the individual words contribute their usual meanings to concatenated meanings which express the concept "change".

But perhaps we were merely unlucky? Perhaps we chanced upon a concept which looked elemental but actually turned out to be complex. Maybe the real primitives are "become", "be", "cease", "different", "same", etc. Let's dig into that possibility.

What does it mean to "become X", where X is an adjective? The meaning can be variously expressed:

(become X, come to be X, get to be X, get X, turn X, grow X, assume the characteristic X)

That's a discouraging number of ways for a "primitive" to be re-expressible -- though if we choose to regard "come to be" and "get to be" as idiomatic concatenations of words, only one of the alternatives makes use of other concepts to explain the one at hand.

As for "different", it implies a whole underlying anecdote about somebody making a comparison, after first making a judgment about relevant things to compare. In the combination of the two concepts -- "become different" --, we furthermore drop mention of the objects being compared. It's simply understood that they are certain attributes of the subject at two points in time.

It is tempting to invent ad-hoc "transformational" explanations for these phenomena. One might conjecture, for example, that "The man changed." is a surface realization of four underlying sentences:

(Man be X at time m. Man be Y at time n. X not equal Y. Time n greater-than time m.)

The trouble with explanations of this sort -- apart from the fact that they introduce growing complexity into the understanding of straightforward utterances -- is that they assign arbitrary primacy to some concepts at the expense of others. Why should

"time n greater-than time m"

be an assumed primitive? May we not equally well conjecture that "time n greater-than time m" is a surface realization of these?:

(Time be m. Time change. Then time be n.)

For that matter, why not view

"Time elapsed."

as a surface form of this?:

"At least one thing in the universe changed."

After all, what is "time" but a nominalized way of talking about the presence and partitioning of change?

The difficulty, it would seem, lies in the very notion of context-independent "conceptual primitives". The metaphor itself is at fault: it calls to mind a fixed set of elements, like those of which matter is composed, out of which all ideas must be compounded. But where concepts are concerned, primitivity is a matter of focus. Shift the perspective a little, and new elements swim into view as fundamentals, while former simples become complex.

A more promising metaphor is the analogy to a vector space. A set of basis vectors is, in a way, a set of "primitives" out of which all the entities in the space can be composed. These primitives have the appealing property that they are only primitive relative to one frame of reference. Rotate your point of view, and what used to come natural as basis vectors are now at an angle; they become easier to express as sums of vectors that lie along new axes. That bears a resemblance to what we have seen in the case of lexical "primitives".

Thus far and no further may the analogy be pushed, however. The elements which span "conceptual space" can be no such uniform set of objects as those in a vector space, while the rules of composition are coextensive with grammar -- at a minimum. Composition of concepts itself contributes to the meaning. (For that matter, it is arguable whether concepts are sufficiently separable to model them as discrete objects at all -- whether simple or composite.) Moreover as "conceptual space" must encompass all things thinkable, the rules of composition must themselves be part of the space. That is, the operators as much as the things operated upon lie within the space to be spanned.

A seeming counterexample to these remarks may be found in the "primitive ACT's" of conceptual dependency theory, as propounded by Schank, Goldman, Rieger, and Riesbeck ([2], [7], [8], [9]). On a close reading, however, the "primitive ACT's" turn out to be *verb paradigms* -- powerful, semantically motivated *generalizations* about large classes of verbs. The names of these paradigms replace specific verbs as building blocks in the "conceptual" representation of an utterance. The effect is to provide strong guidelines for the inference of unstated information, for the comparison of related utterances, for paraphrasal, etc.

To represent a particular verb in terms of these ACT's, however, it is necessary to augment each ACT with various substructures which detail the manner, the means, the type of actor or object, etc. No reduced set of representatives is as yet offered for the adverbs, nouns, adjectives, etc. in terms of which the "primitive ACT's" are qualified. If such additional condensation were attempted, the elaboration of a given utterance in terms of the *full* set of "primitives" might well ramify without practical end. In other words, reduction of the set of names for nodes (and labels for arcs) must be purchased at the expense of extending the number of them required to represent each utterance.

In conceptual dependency representation, just as in the semantic networks" of Quillian [6], Simmons ([11], [12]), Slocum, and others, reality ultimately appears as a shimmering web, every part of which trembles when any part of it is touched upon. Taken in its totality, the system -- as yet -- is entirely compatible with skepticism about a *comprehensive* set of "conceptual primitives"

In any case, the verbal "senses" proposed here lie at a far lower level of generality than the "primitive ACT's" used in conceptual dependency theory. In terms of that theory, they come closest to the so-called "CONCEXICON entries" used by Goldman in realizing surface expressions of a concept from its conceptual representation [2]. Given a primitive ACT, Goldman narrows it down to a particular "CONCEXICON" entry by applying the tests in a discrimination tree to the rest of the structure in which the ACT appears.

Our lexical "senses", therefore, are left with a humbled role. If they span anything, it might best be thought of as "communication space", not "conceptual space". Even in this light, they are a hugely redundant basis, and a not at all unique one. They form no inventory of the experiences being communicated about; "meaning" is still a step removed, still evoked rather than embodied by the elements of this basis.

If we persist in calling these things "senses", it is because that is the traditional term for what is *brought to mind* as the synonym sets of a given word are enumerated. The tie-in with meaning is tenuous, but the human user is able to supply it. There is at least this much justification for the term: synonym sets, more forcefully than words, direct attention to the points at which a tie-in must be made between the tokens of communication and the underlying representation of "world knowledge"

In a full-fledged system for processing natural language, then, we must envision the "dictionary of senses" as a component stretching vertically across the "upper" layers. Its "sense data items" must link, in some way, to the deeper-lying data structures which encode "knowledge of the world" (the "pragmatic component"). The "key data items" and "phrase data items" register tokens to be expected or employed in "surface" utterances. Global and local attributes recorded in the various data items guide parsing and interpretation. Where one takes it from there depends upon the linguistic approach to be used.

References:

[1] Feigenbaum, Edward A. (1963), "Simulation of Verbal Learning Behavior", in *Computers and Thought*, eds. E. A. Feigenbaum and J. Feldman, McGraw Hill.

[2] Goldman, Neil (1975), "Sentence Paraphrasing from a Conceptual Base", *Communications of the ACM*, February, 1975, Vol. 18 No. 2.

[3] Harris, Larry R. (1972), "A Model for Adaptive Problem Solving Applied to Natural Language Acquisition", Cornell University, Ithaca, N.Y. PB-211 378.

[4] Kelly, Edward, and Stone, Philip (1975), "Computer Recognition of English Word Senses", Chapter IV, North-Holland Publishing Co, Amsterdam.

[5] Lamb, Sydney M., and Jacobsen, William H., Jr. (1966), "A High-Speed Large-Capacity Dictionary System", in *Readings in Automatic Language Processing*, ed. David G. Hays, American Elsevier Publishing Company, New York.

[6] Quillian, M. Ross (1968), "Semantic Memory", in *Semantic Information Processing*, ed. Marvin Minsky, The MIT Press, Cambridge, Massachusetts.

[7] Schank, R., Goldman, N., Rieger, C., and Riesbeck, C. (1973), "Margie: Memory, Analysis, Response Generation, and Inference on English", *Proceedings, Third International Joint Conference on Artificial Intelligence*, Stanford Research Institute, Stanford, California.

[8] Schank, Roger C. (1973), "Identification of Conceptualizations Underlying Natural Language", in *Computer Models of Thought and Language*, eds. R. Schank and K Colby, W. H. Freeman & Co., San Francisco.

[9] Schank, Roger C. (1973), "The Conceptual Analysis of Natural Language", in *Natural Language Processing*, ed. Randall Rustin, Algorithmics Press, Inc., New York.

[10] Schmidt, Charles T. (1970), "A Dictionary Structure for Use with an English Language Preprocessor to a Computerized Information Retrieval System", Naval Postgraduate School, Monterey, California. AD 710 363.

[11] Simmons, R. F., and Slocum, J. (1972), "Generating English Discourse from Semantic Networks", *Communications of the ACM,* October 1972, Vol. 15 No. 10.

[12] Simmons, R.F. (1973), "Semantic Networks: Their Computation and Use for Understanding English Sentences", in *Computer Models of Thought and Language*, eds. R. Schank and K. Colby, W. H. Freeman & Co., San Francisco.

CURRENT BIBLIOGRAPHY

Despite repeated predictions to the contrary, both the selection of material for this issue and the choice of subject categories are tentative. The Editor and his collaborators have found the reconstruction of intellectual and mechanical systems more onerous than they had expected.

Completeness of coverage, especially for reports circulated privately, depends on the cooperation of authors. Summaries or articles to be summarized should be sent to the editorial office, Twin Willows, Wanakah, New York 14075.

Many summaries are authors' abstracts, sometimes edited for clarity, brevity, or completeness. Where possible, an informative summary is provided.

The Informatheque de linguistique de l'Universite d'Ottawa, Dermot Ronan F. Collis, Director, provides a portion of our entries. AJCL gratefully acknowledges the assistance of J. Beck, B. Harris, and D. Castonguay.

See the following frame for a list of subject headings with frame numbers.

SUBJECT HEADINGS

GENERAL 26
Chinese 35
PHONETICS - PHONOLOGY . . 36
Recognition 38
  Chinese 43
WRITING 44
Recognition 44
  Chinese 45
Synthesis
  Chinese 46
Text Input
  Chinese 48
Character sets
  Chinese 50
Chinese 54
LEXICOGRAPHY - LEXICOLOGY
Statistics 56
Text Handling 57
Dialectology 58
Thesauri 58
GRAMMAR 59
Generator 60
SEMANTICS - DISCOURSE . . 60
Memory
  Question Answering . . 67
Text Grammar 67
LINGUISTICS
Methods 68
  Mathematical 69
COMPUTATION . . . .
Programming 70
  Languages 71
Pictorial Systems 72
DOCUMENTATION 73
Classification 76
Retrieval 76
TRANSLATION 76
SOCIAL-BEHAVIORAL SCIENCE 79
Anthropology 82
Psychology 83
  Learning 84
HUMANITIES 85
Concordance 86
Analysis 86
INSTRUCTION 88
BRAIN THEORY 91
ROBOTICS

Putnam and Clarke and Mind and Body

Yorick Wilkes
Artificial Intelligence Laboratory, Stanford

British Journal for the Philosophy of Science 26:213-225, September 1975 ISSN 0007-0882

Putnam argues for a satirical privacy for machines by asserting that, just as it makes no sense to ask John how he knows that he is in pain, so it makes no sense to ask a Turing Machine (TM) how it knows that it is in state *A*. When addressed to an abstract TM the question is absurd, but not when addressed to a physically realized TM. Putnam equivocates about the notion of state, discussing only abstract TM's when introducing the notion of state, but making an argument which is coherent only with respect ot a physically realized TM. There is thus, in effect, a confusion between 'state' as of an automaton and 'state' as of a real machine which is executing a program which realizes that automaton--any of a number of states of the machine might correspond to one state of the automaton embodied in the program. Thus Putnam's argument fails. Clarke's criticisms of Putnam are misguided but instructive. A more serious notion of machine privacy can be constructed by noting that it is impossible to inter the machine's real activitity determinately from the content of registers.

A Graphical Programming System with Speech Input

Chacko C. Neroth
University of California, Berkeley

Computers & Graphics 1:227-231, 1975

The experimental problem solving environment is one of formulating specifying, debugging and executing (algebraic) procedures interactively on a small processor. The speech recognition system is a real time, syntax directed, limited vocabulary, highly cost effective scheme specifically tailored to this environment. The data transformation operations of the language are verbally specified and the control flow is specified graphically as a two-dimensional directed graph. The semantics of the latter structure is independent of the time sequence of its input. An input restricted (conditional input) pseudo-finite state machine model is used for the continuous syntax checking of the input on an atomic token basis and for directing the speech recognizer.

Computerized Natural Language Information System

Stewart N. T. Shen
Computer Science Department, Virginia Polytechnic Institute and State University, Blacksburg

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 573-588

General problems in NL processing are discussed and a methodology is presented. A NL system should consist of a supervisory module which reads and interprets certain input sentences stated in some specific way. These sentences tell the system what kind of job is being done. The system would have various syntactic, semantic, and pragmatic processing modules available to it. Technological developments may well make it practical for individuals to have CNLIS (Computerized Natural Language Information System) terminals in their homes. A typical user terminal may include a microcomputer an interactive TV, and an electric typewriter. In the computerization of Chinese, a simplification of the written characters is urged.

GENERAL

Design Concepts of Chinese Language Data Processing Systems

Yaohan Chu, Chu
Department of Computer Science, University of Maryland, College Park

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 117-136

Five types of Chinese language data processing systems are discussed. 1) Accept assembly code in English and hand-coded Chinese data and use an expanded subroutine library. 2) Accept assembly code and data, both in Chinese, by adding a pre-assembler and translators to the manufacturer-supplied assembler and linkage editor. 3) Accept a high-level Chinese programming language and Chinese data. 4) Those which use the Chinese-language-oriented postfix string as the machine language. 5) In which the high-level language itself is the machine language (i.e. one-level language). This type of data processing system has no intermediate language, no assembly language, no relocatable language, and no absolute language.

Design Philosophy of a Chinese-Oriented Computer

John Y. Hsu
Department of Computer Science and Statistics, California Polytechnic State University, San Luis Obispo

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 135-150

Basic design philosophy of a Chinese-oriented computer includes consideration of the idiosyncrasies of such a computer. The following topics are discussed: 1) Internal coding of Chinese characters, 2) Chinese Input/Output devices, 3) Instruction Repertorie to Manipulate Chinese Characters, and 4) Miscellaneous. The design approach toward such a Chinese-oriented computer is also commented on.

GENERAL

Is Technology Ready for Chinese/Japanese Data Processing

B. J. Greenblott, and M. Y. Hsiao
IBM Poughkeepsie, New York

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Acadenia Sinica, 151-161

The technology for the computer processing of Chinese characters on a large scale is almost around the corner. The major bottleneck is the training required to key in 8,000 different Chinese characters quickly and correctly by either a set of keyboards or some cleverly combined coded form or keyboard design. Advances in LSI technology, mechanical or magnetic keys, CRT, etc., will all contribute to the realization of a data processing system capable of handling ideographic languages. An automatic pattern recognition system was not chosen to represent the major future trend because its technical development is still beyond the level of practical large scale implementation.

Conversation, Cognition and Learning

Gordon Pask
System Research Ltd., Richmond, Surrey, England

Elsevier, Inc., Amsterdam and New York, 1975, $37.00/Dfl 96.00, xii + 570 pages, ISBN 0-444-41193-3

This book describes a theory of man/man or man/machine conversations and cognitive processes (with emphasis upon the dynamics of learning and teaching at an individual level) together with several special experimental methods and practical applications. Most of the illustrations and data supporting the argument stem from education, course design, and similar fields and the material is relevant to epistemology, subject matter organisation, as well as such disciplines as pedagogy, computer aided instruction etc. Some experiments, however, deal with laboratory learning and the acquisition of perceptual motor skills, and an attempt is made to identify the theory and methods with many standard paradigms in social and experimental psychology. An account of consciousness and self-reference is given in the theory.

Information Processing and Cognition: The Loyola Symposium

Robert L. Solso, editor
Loyola University of Chicago

Halsted Press Division, John Wiley & Sons, New York, 1975 ISBN 0-470-81230-3 HC $19.95

Contents

Preface xi

SECTION 1

1 MEMORY, PERCEPTION, AND DECISION IN LETTER IDENTIFICATION, W. K. Estes 3

Introduction 3

Apparent Effects of World Context on Perception of Letters 6

Detection Procedures and Revival of the Redundancy Hypothesis 7

The Postexposure Probe Technique 9

An Interpretation of the Role of Positional Uncertainty 16

Levels of Information Processing 18

A Model for Levels of Information Processing in Letter Identification 19

Interpretations of Empirical Phenomena 22

Discussion 26

References 29

2 STUDIES OF VISUAL INFORMATION PROCESSING IN MAN, M. S. Mayzner . . 31

Introduction 31

Current Hardware System and Associated Software Package 34

General Research Strategy 35

A Brief Review of Sequential Blanking and Displacement 36

Dynamic Visual Movement 39

Dynamic Visual Movement 39

Subjective Color Experiences Associated with Dynamic Visual Movement 42

Pattern Recognition Mechanisms as Found in Overprinting Paradigms 44

Overview 52

References 53

3 ATTENTION AND COGNITIVE CONTROL,
Michael I. Posner and Charles R. R. Snyder 55

Introduction 55

Automatic Pathway Activation 56

A View of Conscious Attention 64

Strategies and Conscious Control 69

The Place of Value in a Judgment of Fact 74

Overview 81

References 82

4 FORM, FORMATION, AND TRANSFORMATION OF INTERNAL
REPRESENTATIONS, Roger N. Shepard 87

Some Central Issues 87

Summary of Some Experimental Findings 96

Theoretical Discussion 103

Acknowledgments and Historical Note 117

References 117

SECTION II

5 RETRIEVAL AS A MEMORY MODIFIER: AN INTERPRETATION OF
NEGATIVE RECENCY AND RELATED PHENOMENA, Robert A. Bjork 123

Negative Recency 124

Related Phenomena 136

Conclusion 142

References 143

6 ENCODING, STORAGE, AND RETRIEVAL OF ITEM INFORMATION, Bennet B Murdock, Jr. and Rita E. Anderson 145

Encoding 153

Storage 158

Retrieval 164

Retrieval at Short and Long Lags 185

References 192

7 WITHIN-INDIVIDUAL DIFFERENCES IN "COGNITIVE" PROCESSES, William F. Battig 195

Some Questions about Current Cognitive Research Practices 197

Processing Differences within Individuals 203

Serial Learning 206

Paired-Associate Learning 211

Verbal-Discrimination Learning 219

Free-Recall Learning 220

General Conclusions 224

References 227

8 CONSCIOUSNESS: RESPECTABLE, USEFUL, AND PROBABLY NECESSARY, George Mandler 229

The Revival of Consciousness 229

Conscious Contents and Processes 236

The Limitation of Conscious Capacity and the Flow of Consciousness 246

Conclusion 252

References 253

DISCUSSION: SECTIONS I AND II 255

SECTION III

9 MEMORY REPRESENTATIONS OF TEXT, Walter Kintsch 269

A Model for Episodic Memory 270

Some Remarks on Text Bases 272

Memory for Text 277

Answering Questions about a Text from Memory 284

Matches at Linguistic and Propositional Levels 286

Postscript: List Learning and Text Learning 291

References 293

*10 COMPUTER SIMULATION OF A LANGUAGE ACQUISTION SYSTEM:
A FIRST REPORT, John R. Anderson 295

Formal Results on Grammar Induction 296

The Role of Semantics 299

Rationale for the Research Approach 301

The Program LAS.1 302

Bracket--The Graph-Deformation Condition 317

An Example of Grammar Induction 324

Assessment of LAS.1 334

Generalizations about Noun Phrases 335

A Prediction about Language Learnability 339

Summary 344

Appendix 345

References 347

11 SEMIOTIC EXTENSION, David McNeill 351

The Study of Performance 352

Production Mechanisms 354

Ontogenesis 356

Evidence for this Argument 361

Further Evidence for the Argument 364

Other Influences on Word Order 365

Emergence of Patterned Speech 367

Semiotic Extension 372

Gestures 373

Conclusion 379

References 379

12 THE CONSTRUCTION AND USE OF REPRESENTATIONS INVOLVING LINEAR ORDER, Tom Trabasso and Christine A. Riley 381

Problem Origins: the Development of Transitive Reasoning 382

Training Results: Serial Position Effects 385

Testing Results: Memory and Inference Correlations 386

Stochastic Retrieval Models 388

What Occurs in Training: The Serial-Position Effect 392

What Occurs in Testing? 394

Reaction-Time Experiments: Six-Term Series Problems 395

Linear Order Is Independent of Input 399

The Use of A Linear-Order Representation 400

Accessing A Linear Order: Strength or Distance? 401

Conclusions 407

References 409

DISCUSSION: SECTION III 411

Author Index 427

Subject Index 433

*This article has been abstracted under SOCIAL-BEHAVIOURAL SCIENCE: PSYCHOLOGY on this fiche.

Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems

Academia Sinica, xiii + 1331 pages

The Symposium was held on August 14-16, 1973 at Taipei, Taiwan, Republic of China. Many of the papers have been abstracted elsewhere on this microfiche.

GENERAL: CHINESE

Interactive Processing of Chinese Characters and Texts

J. T. Tou, J. C. Tsay, and J. K. Yoo
Center for Informatics Research, University of Florida, Gainsville

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica 1-28

The system provides a tool to teach pupils how to write Chinese ideographs, how to make proper pronunciations, and how to translate into a foreign language, and features dynamic display of characters. The system can also perform text-editing operations. Techniques for Chinese character representation, based on chain codes for stroke sequence, and dictionary generation, in which each character of subcharacter is represented as a subroutine in the dictionary, are introduced. Text-editing routines are discussed and the paper concludes with an illustrations of text-editing operations. The final edited text can be transcribed from the display scope for making hard copies. The system will be further developed for editing maps, for typesetting and for use as a Chinese typewriter.

A Small Computer in the Phonetics Laboratory

Claes-Christian Elert
Professor of Phonetics, Umea University, Sweden

World Papers in Phonetics, The Phonetic Society of Japan, Tokyo, 145-162, 1974

With adequate programming facilities at hand the phonetician would be able to make his table-top computer perform pratically everything that was done earlier by conventional equipment for analysis and registration. In addition, data may be stored for automatic processing, and sequences of events, such as qualitative or quantitative variations of parameters or stimuli in experiments with human subjects, can be governed according to a pre-set program, or by incoming signals of random pulses. Topics considered: 1) the nature of available equipment, 2) programming, 3) speech analysis, 4) speech synthesis, 5) the computer in experimental work, 6) dialectology and phonology, 7) teaching phonetics.

PHONETICS-PHONOLOGY

A Study of Time-Domain Speech Compression by Means of a New Analog Speech Processor

I. M. Bennett, and J. G. Linvill
Department of Electrical Engineering, Stanford University, Stanford, California 94305

Journal of the Audio Engineering Society 23:713-721, November 1975

Time-domain speech compression using the SDA (sample, discard, about) procedure at compression ratios of 0.25 to 0.75 is studied by means of a new analog speech processor and minicomputer algorithms. Fourier transform methods have been used to establish a correspondence between the quality of the reconstructed compressed speech waveforms and the subjective recognition of compressed speech. The result of two psychoacoustic experiments indicate that 1) the interruption frequency should be equal to the pitch frequency of the voice waveform for optimum recognition of the compressed speech, and 2) smoothing of the discontinuities with electronic techniques significantly improves the recognition of the compressed speech. The optimum smoothing parameters, window width and characteristic function, are also obtained from this study.

On the Characteristics of Individual Vowels and the Statistical Characteristics of Formant Frequency Patterns in Connected Speech

Yoshinari Kanamori
Research Institute of Electrical Communication, Tohoku University, Sendai 980, Japan

Systems - Computers - Controls, 6, No. 1:22-30, 1975

The loci of formant frequency patterns of vowels in many kinds of CVC contexts were represented in F_1-F_2 space. The areas enclosing these loci were obtained for each vowel. The positions of vowels in connected speech lie inside an area surrounded by the isolated vowels because of the neutralization of vowels in connected speech. The faster the speaking ra e, the more the areas tend to concentrate deeper inside. Also, the areas of individual vowels overlap each other and the faster the speaking rate, the more the overlapping areas increase. The overlapping areas were estimated in F_1-F_2 space and, to investigate the effect of F_3 in F_1-F_2-F_3 space. The distribution of F_3 is nearly approximated by the normal density function, because the effect on imbalance of the vowel occurrence frequencies is not clearly observed in the frequency distribution of F_3. Areas reflecting the bound of articulatory movement in the acoustic domain were obtained from the loci of formant frequencies represented in F_1-F_2 and F_1-F_3 spaces. We conclude with a comparison of the discussed areas and those obtained from the articulatory model by Lindblom.

PHONETICS-PHONOLOGY

Epoch Extraction of Voiced Speech

T. V. Anathapadmanabha, and B. Yegnanarayana
Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India

IEEE Transactions on Acoustics, Speech, and Signal Processing 23:562-570, December 1975

A general theory of epoch extraction of overlapping nonidentical waveforms is presented and applied to outputs of models of voiced speech production (model 1, impulse excitation of a two-resonator system; model 2, glottal wave excitation of a two-resonator system) and to actual speech data. Some typical glottal waveshapes are considered to explain their effect of the speech output. The points of excitation of the vocal tract can be precisely identified for continuous speech and it is possible to obtain accurate pitch information by this method even for high-pitched sounds.

Real-Time Digital Hardware Pitch Detector

Ronald W. Schafer
Department of Electrical Engineering, Georgia Institute of Technology, Atlanta 30332

John J. Dubnowski, and Lawrence R. Rabiner
Bell Laboratories, Murray Hill, New Jersey 07974

IEEE Transactions on Acoustics, Speech, and Signal Processing 24:2-8, February 1976

A high-quality pitch detector has been built in digital hardware and operates in real time at a 10 kHz sampling rate. The hardware is capable of providing energy as well as pitch-period estimates. The pitch and energy computations are performed 100 times/s (i.e., once per 10 ms interval). The algorithm to estimate the pitch period uses center clipping, infinite peak clipping, and a simplified autocorrelation analysis. The analysis is performed on a 300 sample section of speech which is both center clipped and infinite peak clipped, yielding a three-level speech signal where the levels are -1, 0, and +1 depending on the relation of the original speech ample to the clipping threshold. Thus computation of the autocorretation function of the clipped speech is easily implemented in digital hardware using simple combinatorial logic, i.e., an up-down counter can be used to compute each correlation point.

A Comparison of Three Methods of Extracting Resonance Information from Predictor-Coeficient Coded Speech

Randall L. Christensen
Naval Weapons Center, China Lake, California 93555

William J. Strong, and E. Paul Palmer
Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602

IEEE Transactions on Acoustics, Speech, and Signal Processing 24:8-14, February 1976

The methods: finding roots of the polynomial in the denominator of the transfer function using Newton iteration, picking peaks in the spectrum of the transfer function, and picking peaks in the negative of the second derivative of the spectrum. A relationship was found between the bandwidth of a resonance and the magnitude of the second derivative peak. Data, accumulated from a total of about two minutes of running speech from both female and male talkers, are presented illustrating the relative effectiveness of each method in locating resonances. The second-derivative method was shown to locate about 98 percent of the significant resonances while the simple peak-picking method located about 85 percent

A Method for the Correctio of Garbled Words Based on the Levenshtein Metric

Teruo Okuda
Systems Design Section, Systems Development Department, Fujitsu Limited, Kawasaki, Japan

Eiichi Tanaka, and Tamotsu Kasai
Department of Electrical Engineering, Faculty of Engineering, University of Osaka Prefecture, Sakai, Japan

IEEE Transactions of Computers 25:172-178, February 1976

Using a method for correcting garbled words based on Levenshtein distance and weighted Levenshtein distance we can correct substitution errors, insertion errors, and delection errors. According to the results of computer simulation on nearly 1000 high occurrence English words, higher error correcting rates can be achieved by this method than any other method tried to date. Short words remain a problem; solving it will probably require utilization of contextual information. Hardware realization of the method is possible, though complicated.

Speaker-Identifying Features Based on Formant Tracks

Ursula G. Goldstein
Department of Electrical Engineering and Computer Science and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, 02139

Journal of the Acoustic Society of America 59:176-182, January 1976

The formant structure of three dipthongs, four tense vowels, and three retroflex sounds was examined in detail for possible speaker-identifying features. Formant tracks were computed for each sound under investigation using covariance-type pitch-asynchronous linear prediction together with a root-finding algorithm. The interspeaker variability of about 200 measurements made on these formant tracks was compared initially with intraspeaker variability through the calculation of F ratios. Those with average F ratios greater than 60 were evaluated further with a probability-of-error criterion. Features that are potentially most effective in identifying speakers are the minimum second-formant value in [ar], the maximum first-formant value in [ar], the maximum second-formant values of [o] and [1], and the minimum third-formant value of [ ] The individual differences apparent in these sounds presumably depend more on speaker habits than on vocal-tract anatomy. The error bound predicted for a speaker identification procedure based on these five features in 0.24%. An identification experiment using only the best two features gave 12 errors out of 80 identifications.

Linear Estimation of Nonstationary Signals

Louis A. Liporace
Institute for Defense Analysis, Communications Research Division, Princeton, New Jersey 08540

Journal of the Acoustic Society of America 58:1288-, December 1975

Implicit in the use of linear prediction is the assumption that within each analysis frame the signal is stationary. The acoustic signal is assumed to be suitably approximated by a recursion which describes a linear time-invariant acoustic system composed of a concatenation of equal-length, constant-diameter nondissipative tubes. That is, assoicated with the coefficients (c) in the recursion is a stylized articulatory configuration which remains fixed throughout the analysis interval. If we allow the coefficients to be functions of time rather than constants we can obtain a more realistic model in which the parameters of the model change continuously and automatically which articulation, rather than discontinuously at fixed intervals. The time-varying area function can be estimated by adapting Wakita's procedure.

A Semiautomatic Pitch Detector (SAPD)

Carol A. McGonegal, Lawrence R. Rabiner, and Aaron E. Rosenberg
Bell Laboratories, Murray Hill, New Jersey 07974

IEEE Transactions on Acoustics, Speech, and Signal Processing 23:570-574, December 1975

The determination of an utterance's pitch contour utilizes simultaneous display (on a 10 ms section-by-section basis) of the low-pass filtered waveform, the autocorrelation of a 400-point segment of the wideband recording. For each of the separate displays (i.e., waveform, autocorrelation, and cepstrum) an independent estimate of the pitch period is made on an interactive basis with the computer, and the final pitch period decision is made by the user based on results of each of the measurements. Formal tests of the method were made in which four people were asked to use the method on three different utterances, and their results were then compared. During voiced regions, the standard deviation in the value of the pitch period was about 0.5 samples across the four people. The standard deviation of the location of the time at which voiced regions became unvoiced, and vice versa was on the order of a half a section duration, or 5 ms. The major limitation of the proposed method is that it requires about 30 min to analyze 1 s of speech.

A Pitch-Synchronous Digital Feature Extraction System for Phonemic Recognition of Speech

Wolfgang J. Hess
Institut fuer Informationstecknik (Datenverarbeitung), Technische Universitaet Muenchen

IEEE Transactions on Acoustics, Speech, and Signal Processing 24:14-25, February 1976

The system has three portions: pitch extraction, segmentation, formant analysis. The pitch extractor uses an adaptive digital filter in time-domain transforming the speech signal into a signal similar to the glottal waveform. Using the levels of the speech signal and the differenced signal as parameters in time domain, the subsequent segmentation algorithm derives a signal parameter which describes the speed of articulatory movement. From this, the signal is divided into "stationary" and "transitional" segments; one stationary segment is associated to one phoneme. For the formant tracking procedure a subset of the pitch periods is selected by the segmentation algorithm and is transformed into frequency domain. The formant tracking algorithm uses a maximum detection strategy and continuity criteria for adjacent spectra. After this step the total parameter set is offered to an adaptive universal pattern classifier which is trained by selected material before working. For sationary phonemes, the recognition rate is about 85 percent when training material and test material are uttered by the same speaker. The recognition rate is increased to about 90 percent when segmentation results are used.

Analysis of Intonation Signals by Computer Simulation of Pitch-Perception Behavior in Human Listeners

Yukio Takefuta
Ohio State University

SIGLASH Newsletter 8, No. 1:1-8, February 1975

Pauses are used to delimit utterances into segments. Linear regression analysis of pitch patterns allows a 4-way classification of slopes of lines: fast rising, rising, level, falling. These are the 4 Fundamental Pattern Features (FPF). A combination of 2 or 3 (of the 4) FPF's per segment of utterance is a pitch pattern (80 possible). An intonation pattern is a combination of pitch patterns. The position of the highest frequency value in the utterance is important. In comparing 2 utterances, if the high point occurs in different segments the intonations are contrastive even if the pitch patterns are the same. Of the 80 possible pitch patterns, some must be recognized as cardinal patterns and same as cognate patterns to the cardinal patterns. Different sets of rules must be used for the "high" segment and the final segment of the utterance.

Graph-Theoretic Cluster Analysis and Its Application to Speech Recognition

Z. Chen
School of Electrical Engineering, Purdue University, West Lafayette, Indiana

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 225-242

A clustering algorithm is mainly a two stage process: 1) selection of a pairwise similarity measure between every two samples or objects in the data set, 2) the similarity measure is used in a sorting procedure whereby groups of similar samples are extracted. In a graph-theoretic clustering algorithm a graph is constructed for the given data and subgraphs G satisfying certain properties are obtained. The clustering algorithm features a flexible method of edge construction (k-nearest neighbor threshold method), which allows the grouping of samples to be more effective, and the generalized Frisch's labelling algorithm, which detects and removes the possible chaining effect in the data. The algorithm is applied to the recognition of nasal consonants.

A Comparison of Several Speech-Spectra Classification Nethods

H. F. Silverman, and N. R. Dixon
Speech Processing Group, Computer Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, New York

IBM Research Report 5584, 18 August 1975

Two measures of performance of speech spectra classification--accuracy and stability--were derived through the use of an automatic performance evaluation system. Over 3000 hand-labelled spectra were used. The most successful classification method involved a linearly-mean-corrected minimum distance measure, on a 40-point spectral representation with a square (or cube) norm. Straight minimum distance is the worst performer. The question of appropriate point representation is really one of adequate information retention. The 80-point representation contains too many components above 3kHz while 20- and 10-point representations contain insufficient information relative to the classes to be discriminated. The value of the norm exponent primarily relates to the weight given extrema in the norm kernel; a heavier weighting (2 or 3) should be placed on extrema.

Speech Recognition and Chinese Voice Input for Computer

Kung-Pu Li
TRW Systems Group, One Space Park, Redondo Beach, California

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 211-223

The machine recognition of Mandarin mono-syllables seems to be feasible at the present. An integrated recognition procedure of monosyllable utterances has also been suggested, and some results are described. The basic syllable structure contains three major parts: initial, tone, and final. The initial contains only consonantal phonemes of four different categories: sonorant, plosive, fricative, aspirate. There are four tonemes in Mandarin Chinese; the pitch contours cover only the final part of the syllable. In the vowel part of a syllable, although seven phonemes are sufficient to describe all possible vowels, the final can also contain dipthongs and tripthongs composed of more than one vowel phoneme. An itegrated recognition procedure of monosyllable utterances has been suggested, and some results are reported.

PHONETICS-PHONOLOGY: CHINESE

Chinese Phonemes Analysis and Synthesis

T. Y. Chou, and K. C. Huang
National Chiao Tung University, Hsinchu, Taiwan, Republic of China

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 1227-1241

The system for producing consonants is a noise generator followed by a pole-zero resonator, while that for producing vowels is a quasi-periodic pulse generator with variable period followed by a resonator with three variable poles called formant frequencies, ranging from 0 to 3 KHz. The results of analyzing, by means of sonagrams obtained from ten male voices, show that the 16 vowel phonemes can be classified into two classes as single and compounded vowel sounds. Some of the synthesized single vowels are very monotonic and can be recognized. Others, with third formant frequency slightly greater than 3 KHz are not as clear, due to the fast decaying of high frequency components in the generated pulses. The compounded vowels are also synthesized by a step variation of formant frequency derived from its components. The result is also well recognizable. The sonogram analysis of a continuous Chinese speech shows that every word and its spelling phonemes are quite independent and separable, and are therfore very different from English speech.

Graphemic Synthesis: The Ultimate Solution to the Chinese Input/Output Problem

Carl Leban
The University of Kansas, Lawrence

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 533-552

In design of IO systems for human graphics it is necessary to simulate the *activity* of writing and not the graphic result of that activity. Orthographic rules are essentially sets of criteria for determining proper *serial order* of graphic signs, upon which further subsets of phonetic, semantic, graphic and other conventions are imposed. Chinese orthography is of the polyalternating polyvariable type in which a number of undefined subsets of graphic signs combine with each other in any of several possible juxtapositional modes according to rules not yet fully elucidated. But if the logography is not converted to an invariable series it cannot be input to, manipulated in, or output from a digital computer. The temporal series in which elements are composed into logographs is variably serial, and therefore computer compatible. Graphemic synthesis is a procedure by which logographs are mechanically produced in a manner simulating the normal writing procedure. Since logographs are synthesized from a small finite set of component elements, there is no need to prestore logographs, but only the small grapheme set. Output in normal logography is achieved as needed only at the output end by reversal of the synthesizing process. It is possible to achieve a synthesis at somewhat less than the ideal level, *pseudographemic synthesis*, and this has been implemented in the SINCO system.

WRITING: RECOGNITION

Feature Extraction on a Finite Set of Binary Patterns

Paul P. Wang, and William S. Hodgkiss, Jr.
Department of Electrical Engineering, Duke University, Durham, North Carolina

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 183-194

A "best" subset of mutually orthogonal features which are minimum in number but sufficient to discriminate a finite set of patterns is chosen from a much larger set of available features in a systematic and deterministic manner by a heuristic program based on the criterion of maximum separability. The unique code words for the finite set of binary patterns are established through a learning procedure derived from a theorem on necessary and sufficient conditions for mutual independence of these vectors over a binary field.

An Experimental System for the Recognition of Handwritten Chinese Characters

Shi-Kuo Chang
Institute of Mathematics, Academic Sinica, Nankang, Taipei, Republic of China

Der Her Lo
Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, Republic of China

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 257-267.

A Chinese character can be thought of as composed from a set of straight line segments. A stroke is ideally a line segment or a concatenation of several straight line segments. Each straight line segment has its starting point, direction and length. There is also a specified sequence among these segments. The specified sequence of line segments for a character is the same as the sequence of their starting points. Therefore, when a character is drawn on the tablet of the digitizer, the output paper tape containing the (x,y) coordinates of sampling points of each of its line segments presents these points in the proper order. The preprocessor produces a sequence of simplified straight line segments, each with a direction code and length, from the paper tape input, and sends the results to the classifier which constructs a dictionary of characters which it then uses in the recognition process. The program achieved 95% recognition for a test sample of 300 Chinese characters.

WRITING: RECOGNITION: CHINESE

Computer-Aided Chinese Character Recognition by Forward Markovian Dynamic Programming

Yung-Lung Ma
Department of Electrical Engineering, National Taiwan University, Taipei, Republic of China

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 269-286

A Chinese character may be called a kind of block picture. Each stroke seems to be a hieroglyph. Curves formed by several strokes or by a continous stroke do not happen very often. Some characters only differ by a single stroke. If the Markovian processes mentioned in this paper are used, detailed recognition for each row and column is available and even a single stroke would not be missed absolutely, so the increasing the degree of correct recognition is by no means a problem. A plane block picture may be divided into plane blocks while a solid one can be discussed by dividing it into solid blocks. The greater the number of layers, the higher the degree of correct recognition. The input pattern is divided into 20 layers for individual recognition. If every layer satisfied the condition, then recognition is complete. If one of the layers has a great difference, then the pattern should be another picture.

The Topological Analysis, Classification, and Encoding of Chinese Characters for Digital Computer Interfacing--Part I

Paul P. Wang
Department of Electrical Engineering, Duke University, Durham, North Carolina

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 417-439

A set of features believed to be useful in classification and recognition and which is deduced from topological properties and heuristic properties is proposed. An encoding scheme offers a unique code word for each character (signature) of a dictionary of about 6,000 items. A three-stage machine recognition system, based upon the optimal multiple category classification principle, has been proposed to solve the problem of automatic reading of Chinese characters. A by product of this research is the development of a topologically based lexicographical ordering for a useful Chinese dictionary. Finally, some recommendations concerning machine recognition of printed Chinese ideographs are made.

WRITING: SYNTHESIS: CHINESE

Software Method in Kanji Information Processing

Teiji Kakinuma
Fujitsu Limited, Minato-ku, Tokyo, Japan

Kenmi Tsukatani
PETY Limited, Chiyoda-ku, Tokyo, Japan

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 983-998

In the FCL (FACOM Composition Language) System, information is punched on paper tape with a Kanji keyboard. The layout data turns out the forms of final printed matter as parameters, punched with an alphanumeric keyboard, and these are applied to the FCL as an input together with the text data. The results of the editing by the FCL are output to a cassette magnetic tape, transferred to the photo type-autosetter, and are printed on film. At this stage a print for correction is completed and this print is placed in the correction processing cycle. The correction processing generates the corrected data concerning errors in the test and layout data and this corrected data is again input to the FCL. The FCL saved file is utilized as the objective of the correction processing. The FRAME program is the portion of the layout control system which defines paragraph groups which have the same character and shape.

Photo-Electrostatic Kanji Printer

Atsushi Ishi, Yoichi Hagiwara, Woshimitsu Masui, and Yoshiyuki Aida
Fujitsu Ltd., Minato-Ru, Tokyo, Japan

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 969-981

This Kanji printer consists of a character generator and a printer. The character generator has a small rotating image disc with 5,376 characters printed on it, and the character patterns are converted into video signals by a vidicon. The printer has an optical fiber tube and a photo-electrostatic recording element. Reproduced character patterns on the surface of the optical fiber tube are recorded on the dielectric coated paper by a photo-electrostatic element. This Kanji printer is capable of printing 100 characters a second, and is usable for any application of printing in Kanji.

WRITING: SYNTHESIS: CHINESE

Designing Storage/Output Units for Chinese Input/Output Digital Computers

Kai Huang
Department of Electrical Engineering, University of Miami, Coral Gables, Florida

S.Gould., Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 931-942

The storage unit (SU) provides a permanent filing cabinet for storing Chinese characters in any predetermined binary-coded form. The stored information is addressable and readable from external control. The SU contains a large-scale cellular array of Read-Only Memories (ROM) with the associated address decoder, control logic, memory address/data registers and sense amplifier to enable readability. The output unit is used for printing or displaying decoded Chinese in ideographic form. It consists of a k2-segment character decoder, two buffer registers, and auxiliary display terminals (DT). The DT may take many forms, such as a multiple-head printer, D/A converter and storage CRT monitor or a graphical display console make with an array of light emitting devices (LED).

Computer-Aided Design of Chinese Character Patterns

Hideo Hirahara, Kiyoshi Kibuchi, and Masamitsu Satou
Informations Systems Research Laboratory, Toshiba Research and Development Center, Kawasaki-City, Japan

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 909-930

The System generates dot patterns from any original design pattern or handwritten items. The source patterns are scanned with a vidicon camera and recorded on magnetic tape. The following procedures are used: 1) Noise elimination and smoothing of scanned patterns, 2) Line enhancement, 3) Matrix size compression, 4) Interactive refinement, 5) Automatic generation of Chinese character read only memory (ROM) patterns. The obtained dot patterns are then translated into a paper tape for a numerical controller which in turn drives the wiring system for the read only memory. Chinese character line printers, CRT displays, and other Chinese character output devices can be implemented by this Chinese character read only memory.

WRITING: TEXT INPUT: CHINESE

A System Design for the Input of Chinese Characters through the Use of Phonetic and Orthographic Symbols.

H. C. Li, S. P. Hu, C. L. Jen, H. Chou, S. Shan, and E. T. Chen
Department of Economics, Bryant College, Smithfield, R.I.

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 501-511

The following principles have been observed in system design: 1) Easy to Learn and Use, 2) Inexpensive to Implement, 3) Higher Input Rate, 4) Unique Code for Dictionary Search, 5) Facilitate Other Related Applications. The input of Chinese characters is through the use of phonetic and orthographical symbols. The total number of symbols needed to transcribe a single Chinese character varies from a lower limit of three to a maximum of eight. A single Chinese character requires a maximum of three phonetic symbols and one intonation notation to indicate pronunciation. By coupling one to four of fifteen orthographical symbols with the pronunication symbols, each Chinese character can be uniquely transcribed into a set of symbols which indicates the pronunciation as well as the orthographical structure of the ideography. No new hardware is required for implementation. With some minor modifications, the keypunch machines and other typewriter-like input peripherals now available on the market can be used immediately.

A New Approach to a Chinese (Tele) Typewriter, Which Can be Used as a Telex, Data Terminal and Computer Input/Output Device

Ye-San Liu
Directorate General of Telecommunications, Taipei, Republic of China

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 489-499

A typing Keyboard is proposed which is arranged like an English teletypewriter, using an 8-unit code as in an ASCII code with even parity Switching a lever key, you will be able to type Chinese or English. When you type Chinese, you need only to push a key three times at most to complete the selection of a character. Then the character will come out by pushing the space bar once. If it is so arranged as to push a key three times for all characters, we shall be able to save the process of pushing the space bar for every character. The rules of decomposing Chinese characters are studied. The key layout of the keyboard is so arranged as to make recognition of key position easy. Four simple typing rules have been determined. There are only 21, out of the 3,000 characters which often appear in current newspapers, that share the same 3 keyed codes, and so they have been treated as exceptions.

PEACE--A Phonetic Encoding and Chinese Editing System

Shi-Kuo Chang, Chi-Shion Chiu, Ming-Hwei Yang, and Bao-Shuh Lin
Computation Laboratory, Institute of Mathematics, Academia Sinica, Nankang, Republic of China

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 29-47

Different Chinese characters may have the same phonetic transcription (using Chinese National Phonetic Symbols), requiring methods to disambiguate homonyms. If however, the Chinese text is coded into *phrases* separated by delimiters, then the *phrases* can often be decoded unambiguously to obtain the corresponding string of Chinese characters. The file structure for the PEACE system consists of a character file and a phrase file. Chinese characters are stored in the form of a composition rule. The phonetic encoding method has also proved to be quite satisfactory, especially for the generation and editing of Chinese texts, where more than 60% of the characters are embeded in phrases. Character encoding rates of the order of 30 characters per minute can be attained with this system.

A New Alphameric Code for Chinese Ideographs

Nelson Ling-Sun Chou
East Asia Library, Rutgers University

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 471-488

The Following are assumed: 1) All Chinese ideographs are composed of one or more components, and thus may be classified by the pattern of these components. 2) Each of the components is in turn composed of one or more graphic elements. The total number of graphic elements is fairly limited. Ideographs may be divided into four major patterns: Horizontal, Vertical, Bordered, and Independent. After identification of an ideograph's pattern, a component's structure, and the basic elements, one can then perform the coding by following rules for: 1) Ideograph as a whole, 2) Components, 3) Elements, 4) Relationship or separation signs, 5) Coding sequence, 6) Component of bordered pattern, 7) Independent ideographs of components.

Chinese Input-Output with Standard IBM Selectric Typewriter Terminal

Ching C. Tsao
IBM Corporation, Armonk, New York

Emerson W. Pugh
IBM T.J. Watson Research Center, Yorktown Heights, New Yor,

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 459-469

A multicorner indexing system has been developed for entering Chinese characters, using the numeric keys of a standard computer terminal. Each character is uniquely coded with a sequence of one to nine digits in a nine-corner extension of Wang's Four Corner System. A 20- by 21- dot array output code has been programmed in APL for use on the IBM 2741 Standard Selectric* Terminal with the fine-plot printing element. These two techniques can be combined to provide an easily learned Chinese input-output system for use on standard computer hardware.

The Creation of a Set of Alphabets for the Chinese Written Language

Kendall L. Su
School Electrical Engineering, Georgia Institute of Technology, Atlanta

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 441-458

A set of symbols has been derived by dividing up each in a set of 4,600 Chinese characters. This set of symbols can represent practically all Chinese characters. These symbols can be used as the alphabets of the language, except that they offer no phonetic information on each character. The spelled-out Chinese is readable without ambiguity and referential aid. The immediate and long-term applications of these alphabets are discussed. Extensive examples are given.

On the Formal Description of Chinese Characters

Marion R. Finley, Jr.
Universite Laval, Quebec, Canada

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 513-531.

Chinese characters maybe defined as objects produced by a generative grammar suitably extended to two dimensions. At the heart of this extension is the coordinate-free configuration operator which, together with its inflections, permits the desired stroke linking to compose a given character. The character so generated corresponds then to a derivation-tree which reveals structural properties common with other characters. This tree in turn leads to a quasi-algebraic expression for the character which can be coded to give character indexing for storage and retrieval purposes. The recognition problem is considered briefly. Some experimental results with a character design language are presented.

The Chiao-Tung Radical System, Part I: The Analysis and Design of the Chiao-Tung Radical System

Ching-chun Hsieh, Yung-wen Hwang, Shu Lin, and Kwei-min Hsu
Computer Science Department, National Chiao-Tung University, Hsinchu, Taiwan, Republic of China

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 49-62

Chinese characters are characterized by strokes which are well distributed in a block space of a definite size. The use of radicals *plus* weight design allows composition of characters from radicals to yield pleasing results. Radicals with fewer strokes are given smaller weights; those with complicated strokes are given greater weights, so that the character obtained from the composition is evenly distributed in the block. The characters thus composed directly look very much like those integral characters which are obtained from dot matrices without decomposition. Using only 496 radicals, 18,713 Chinese characters can be generated. The radical system proposed is a precedence grammar

The Chiao-Tung Radical System, Part II: Character Composition and Methods to Represent Radicals

Ming-Wen Du, Ching-Chun Hseih, Shu-Hong Chi, and Shu-Ching Chu
Computer Science Department, National Chiao-Tung University, Hsinchu, Taiwan Republic of China

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 63-78

There are two steps in composing a Chinese character from its radical formula. 1) Calculate the position of each radical and the area occupied by each in the character. 2) Compress each radical and place the radicals into their right positions. There are seven methods for representing the radical system in a computer: 1) dot matrix method, 2) absolute line segment method (the method adopted in implementing the system), 3) relative line segment method, 4) core method, 5) chain code method (basically a line segment method), 6) analytical method, 7) mixed mode, which requires the smallest memory space (7.992 K Bytes) to store all 496 radicals.

The Upper-Right Corner Indexing System for Chinese Language and a New Chinese Character Encoding System

T. Y. Kiang
Department of Electrical Engineering, National Taiwan University, Taipei, Republic of China

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 79-93

All Chinese characters may be classified into three classes: 1) Characters with no common radicals, 2) Characters composed of a common radical and a main part, 3) Characters including more than one common radical. 100 common radicals were selected for use in the system and nearly 70% of the characters belong to class 2. Since most of the common radicals are on the left sides of characters, we suggest indexing a character according to its upper or right stroke form. In the resulting system it is necessary to learn only 47 indices. The characters as well as the main parts are equally distributed under the indices. The system is easy to learn, fast in operation, has a large vocabulary, low in cost and is readily implementable on a mini-computer.

Character Identification Using the Phonetic Code and the Four-Corner Code (in Chinese)

T. Lee, and C. Yang
National Tsing-Hua University, Hsinchu, Taiwan, Republic of China

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 95-101

我國現有的各型計算機總數已相當不少。其中多數的用途尚限於教學與學術研究，眞正用於公私業務及爲日常生活服務者，爲數不多。就算是有，也不見得已發揮他們應當發揮的效能。造成這個事實的基本原因之一，是到目前爲止計算機界還沒有推出一套操作簡單，訓練人員容易，而價格又相當便宜的中文輸入輸出系統。國內外的科技界，爲了促使這個系統早日出現，經多年來的努力，成果已相當大。單就中國字的編號，常用字次序排列的研究等，就有不少著作 [1-4]。

中國文字是圖形文字，六書中又以形聲字佔大多數。如何利用這兩大特色，把中國字加以整理和編碼，以配合計算機中文輸入和輸出的研究，是目前有志一同者努力的方向和目標之一。作者深以爲傳統的部首筆劃法檢字速度太慢，不適於配合高速機器。在前兩篇已發表的文獻中 [5,6]，主張用注音符號及四角號碼作中文輸入的檢字索引。

Use of Contextual Information in the Design of a Chinese Character Pattern File

I-Lieh Huang
IBM Watson Research Center, Yorktown Heights, New York

S.Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 1007-1012

We assume that Chinese characters are digitized into n by n matrices of 0's and 1's and represented in packed from in computer storage, usually in a magnetic disk. Assume that an unblocked indexed sequential method is used for organizing the c.p. (character pattern) file. Each record contains a 2-byte key and 34n bytes of data where n is the number of Chinese character patterns contained in the record. We assume here that we use 2 bytes for character ID and 32 bytes of character patterns. If we increase the number of character patterns in each record from one to four, we reduce the number of records per track (disc memory) by a factor less than 2 A study of contexts of occurrence will tell us what c.p.'s should be stored together. Thus a record should contain not only the c.p. of a given character, but the c.p.'s of some characters that would follow the given character with high probability.

WRITING: CHINESE

An Intelligent Terminal for Chinese Character Processing

F. F. Fang, C. N. Liu, and D. T. Tang
IBM Thomas J. Watson Research Center, Yorktown Heights, New York

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 103-114

The proposed terminal system consists of three modules: Input, Output, Control. The input module includes: 1) a character board consisting of a position sensing input device or an array of keys which may be used to input a selected character by transforming its x, y positions to any desired character code, 2) a general purpose keyboard which may contain 50 to 300 keys each of which is associated with a discriminating mask, and 3) a set of discriminating masks which, when used singly or as a group, perform the character set selection logic. The output module consists of: 1) a character description file which is organized according to the character code and contains information for generation of characters for printing or display, and 2) a printer or display for physical output of characters. The control for the input and/or output system is obtained through either software or firmware. Various degrees of intelligence may be programmed to achieve additional character selection and character generation features.

Techniques for the Implementation of a Chinese Input/Output System

G. W. Crawford, and S. H. Chung
Westinghouse Electric Corporation, Pittsburgh, Pennsylvania

S. Gould Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 961-968

The input problem is solved using an extension of Wang's *Four-Corner Dictiorary Method* that uniquely describes a Chinese word by a six digit number. The technique is then simplified by utilizing a five digit number and an interactive CRT terminal that allows complete resolution of ambiguities. Various methods of output are briefly considered and it is concluded and a Computer Output Microfilm (COM) unit is the most logical device presently available. A specific example of an implementation of a Chinese input/output system based on a. CDC 7600 computer driving a Stromberg-Datagraphix COM unit is described and a specific example of the output included.

WRITING: CHINESE

The Modern Chinese Computer System

Y-Tze Kuo
B.P. 5080, Kinshasa, Zaire, Africa

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 957-959

Input is handled by the author's mixed typing system Modern Chinese typewriter which has a total of 2400 normal and half-size Chinese words which permit the formation of the necessary 8000 common Chinese words. With this smaller number of Chinese words it is possible to represent them by different punched holes on the same kind of punch card as in the occidental computers. Any occidental computer and any new Chinese computer could be used with the system for Chinese language as well as for English and other languages. Output is through the author's Modern Chinese typewriter equipped with automatic typing and positioning devices.

The Chiao-Tung Radical System, Part III: Keyboard Design and an Implementation of the Chiao-Tung Radical

Yung-Wen Huang, Ching-Chun Hsieh, Sie-Lin Liu, and Tai-Cherng Chen
Computer Science Department, National Chiao-Tung University, Hsinchu, Taiwan

S. Gould, Ed., Proceedings of the First International Symosium on Computers and Chinese Input/Output Systems, Academia Sinica, 943-956

The system consists of: 1) HP 2100A mini-computer, 2) keyboard of 640 keys, 3) or keyboard of 88 keys, 4) tape reader; 5) electrostatic printer, 6) display scope. Chinese characters may be fed into the computer either through keyboards or through paper tape Output is printed by the printer and also displayed on the scope. The input radicals are combined into characters by the composition method in the computer. After three hours of familiarization, an operator may reach a speed of 10 characters per minute. It is estimated that a speed of 30 cpm may be reached in a month. To be compatible with other input methods (phonetic alphabets, four-corner code, standard telephone code, etc.) only a table is needed. Since it takes merely 1 ms to compose a character, time sharing may be utilized.

LEXICOGRAPHY-LEXICOLOGY: STATISTICS

Computer Generated Word Classes and Sentence Structures

A. Tretiakoff
Commissariat a l'Energie Atomique, BP 6, 92, Fontenay aux Roses, France *Information Processing 74, North-Holland Publishing Co., 919-920, 1974*

First a dictionary of the corpus (here is an extract from S. Maugham's *The Painted Veil*) is produced. Classification of words is based on the maximum information principle which considers that for a given number of groups, the greater the quantity of information, the better the distribution of the words into these groups. The words are distributed into two groups in such a way that the quantity of information associated with this classification is maximized. Dichotomization is carried on until the statistical uncertainty on the amount of information is greater than the gain of information obtained by a new dichotomy. For each sentence of the corpus, the code produces a structure based on the degree of correlation of two consecutive words. Inside the sentence consecutive words are connected two by two in order of decreasing degree of correlation.

Text Connexitivity and Word Frequency Distribution

Hans Karlgren
KVAL, Stockholm

Hakan Ringbom, Ed., Style and Text, Sprakforlaget Skriptor, Stockholm, 335-348, 1975
ISBN 91-7282-095-0

The distribution of a word in a text can be described in two ways. by the manner in which it differs from that in some other texts and by the manner in which it varies from one part of the text to another. Thematic words are those which stand out in comparison with a given background. The use of the proper statistical techniques (a measure due to Hassler-Goransson which is defined as the chi-square value divided by the number of degrees of freedom is suggested) makes it possible to study the way in which words enter and leave the scene in a pattern which characterizes the text in much the same way as the *intrat s* and *exit s* determine a play. Using more sophisticated methods it is possible to study the strength of connection between any two parts of the text and a segmentation of the text into internally more strongly connected and mutually more loosely connected portions can be tested or even mechanically suggested.

LEXICOGRAPHY-LEXICOLOGY: TEXT HANDLING

Machine Readable Texts in Latin (with Special Reference to the Didactic Implications of a Computer Aided Method)

E. M. Goldstein
University of Ottawa, Ontario, Canada

SIGLASH Newsletter No. 4:1-3, October 1975

Four items of information are recorded for each entry. (1) the *form*, the word as it is in the text, (2) the *lemma*, the form of the word in a dictionary, (3) a detailed *reference* to the form, (4) its coded *analysis* from morphological and syntactic viewpoints. It was found that for Caesar's *Commentarii de bello gallico* the 704 words which occurred more than 10 times covered 86.03% of the text. Studies of this sort are important in determining what vocabulary should be taught to students; there is no immediate need to teach words of low occurrence.

An Organization for a Dictionary of Senses

Dick H. Fredericksen
IBM Thomas J. Watson Research Center, Yorktown Heights, New York

IBM Research Report 5548, 4 June 1975

"Senses" are represented separately from "wordings" and the mutual connections between them are made explicit in both directions. "Wordings" may be either single words or multi-word phrases which are on the same footing with regard to "sense" connections. Each word is associated with an exhaustive list of the phrases in which it occurs. Classifiers and features, drawn from appropriate sets, may be attributed separately to words, to phrases, to senses, or to particular senses of words or phrases (i.e. to particular wordings of senses). The data items which represent senses are globally chained, and may be exhaustively listed. The data items which represent words are accessible as "leaves" of a lexical tree and may be retrieved either by lookup or volunteered in alphabetical order. The "senses" represented in this dictionary are not a set of primitives into which human experience can be decomposed; meaning is still a step removed, still evoked rather than embodied by the elements of this basis. In a full-fledged system for natural language processing the "dictionary of senses" could be envisioned as a component stretching vertically across the "upper" layers. The "sense data items" must link to the deeper-lying data structures which encode "knowledge of the world."

LEXICOGRAPHY-LEXICOLOGY: THESAURI

The Structuring of an Associative Empirical Thesaurus of English

Christine M. Armstrong, and J. R. Piper
Medical Research Council, Speech and Communication Unit, University of Edinburgh

SIGLASH Newsletter 8, No. 2-3:1-6, April-June 1975

A collection of English language data based on free association is organized as a network: node = word, arc = associative frequency. We obtained 100 responses for each of 8,400 words; the net has 56,000 nodes. For each of 35,000 words we can obtain an environment of related words and this environment is fairly large and relevant in content for 8400 stimulus words. The environment can be clustered into subsets, which turn out to be a semantic sorting of the environment. The search can be forward (stimulus to associate), inverse, or in both directions. Since growing environments can produce ponderously large subsets of network, environments are limited by techniques involving transmittance to each node, the number of paths traversed which lead to each node, path length, and a frequency cut-off. It seems better to consider forward and inverse environments separately.

Optimal Encoding of Linguistic Information

Kazuhiko Ozeki
NHK (Japan Broadcasting Corporation) Technical Research Laboratories, Tokyo, Japan 157

Systems - Computers - Controls 5, No. 3:96-103, 1974

Stochastic context-free languages as models of NL. A sentence is parsed first and transformed into a sentence derivation, and then this derivation, expressed as a string of results of "productions," is coded by coding each production. For this coding procedure, it is shown that if the productions are divided into groups of the same left-hand side and if each production group is Huffman-coded, then the mean code length is less than the sum of the entropy and the mean number of steps of sentence derivations. Futhermore, under certain conditions this code becomes optimal in the sense that the mean code length and the entropy coincide, that is, there is no uniquely decodable code with shorter mean code length.

GRAMMAR

Stochastic Context-Free Grammar and Markov Chain

Kazuhiko Ozeki
NHK Technical Research Laboratories, Tokyo, Japan 157

Systems - Computers - Controls 5, 3:104-110, 1974

A stochastic context-free grammar (scfg) is an automaton which stops when a sentence is generated. However, in investigating the information theoretic properties of a long text is convenient to use a scfg-based system which returns to the start symbol and begins the generation of another sentence as soon as it completes the generation on one sentence. The system becomes a Markov chain having the set of all derivations as the state space, which is called here the Markov chain associated with an scfg It turns out that: 1) The Markov chain assoicated with an scfg is irreducible. 2) For the chain to be recurrent it is necessary and sufficient that the language generated by the scfg be a probability space. 3) For the chain to be positive recurrent, it is necessary and sufficient that the mean number of steps of the sentence derivations be finite. 4) It is well known that when a Markov chain is positive recurrent it has an invariant distribution and its entropy per step H_1 is defined. If an scfg satisfies certain conditions, we have $H_1 = H(E)/(M(E) + 1)$.

On the Generation of English Sentence

Franz Huber
Cardiovascular Pulmonary Research Laboratory, University of Colorado Medical Center, Denver 80220

IEEE Transactions of Computers 25:90-91, January 1976

Beckmann's error-detecting code model for the structure of natural languages is used in a simple efficient procedure for generating a large number of English sentences with simple grammatical structure and a high rate of word repetition--such as diagnostic messages of a compiler, messages in interactive systems, etc. The grammar described is implemented with a state table with 8 states: Article, Numeral, Adjective, Noun, Conjunction, Preposition, Verb, Auxiliary Verb. The syntactic type of the varius words is specified by the location of the entry in the dictionary. Sentences are formed by supplying the routine with a sequence of pointers to dictionary entries; the routine itself checks whether this is a possible sentence, calculates the correct check morphemes, and puts a period at the end.

SEMANTICS-DISCOURSE

Montague Grammar and Transformational Grammar

Barbara Partee
Department of Linguistics, University of Massachusetts, Amherst

Linguistic Inquiry 6: 203-300, Spring 1975

A truth-definition (in the manner of Tarski) or something to the same effect must be a *part* of any adequate semantic theory. The syntax of a Montague grammar is a simultaneous recursive definition of all of the syntactic categories of the language. For every syntactic category there must be a unique corresponding semantic category and for every syntactic rule that combines (operates on) phrases of categories A and B to produce a category C, there must be a unique semantic rule that operates on the corresponding semantic interpretation to give a semantic interpretation for the resulting phrase; that interpretation will be of the semantic category corresponding to the syntactic category C. Techniques involving labelled bracketing and "starred variables" enable the additions of transformations to Montague grammars.

Contemporary Research in Philosophical Logic and Linguistic Semantics

D. J. Hockney, E. William Harper, and B. Freed, eds.
University of Western Ontario

D. Reidel Publishing Company, Dordrecht, Holland, 1975
HC: ISBN 90-277-0511-9 *PC: ISBN 90-277-0512-7*

Contents

Preface vii

Counterfactuals and Comparative Possibility, by David Lewis 1

Presuppositions, by Robert Stalnaker 31

Incomplete Assertion and Belnap Connectives, by Bas C. van Fraasen 43

Dimensions of Truth, by Hans G. Herzberger 71

Speaking of Nothing, by Keith S. Donnellan 93

The Structure of Efficacy, by Zeno Vendler 119

Harris and Chomsky at the Syntax-Semantics Boundary,
by Ray C. Dougherty 137

Some Transformational Extensions of Montague Grammar,
by Barbara Partee 195

Hedges: A Study in the Meaning Criteria and the Logic of
Fuzzy Concepts, by George Lakoff 221

Coments: Lakoff's Fuzzy Propositional Logic, by Bas C. van Fraasen 273

On the Semantics of Negation, by Howard Lasnik 279

Verbs of Bitching, by James McCawley 313

Understanding Language: An Information-Processing Analysis of Speech Perception, Reading, and Psycholinguistics

Dominic W. Massaro, Ed.
Department of Psychology, University of Wisconsin, Madison

Academic Press, New York, 1975. ISBN 0-12-478350-8
HC: $16.50

Contents

List of Contributors ix

Preface xi

Part I Introduction

1 Language and Information Processing, Dominic W Massaro

I. Introduction 3

II. Information Processing 5

III. Auditory Information Processing 7

IV. Visual Information Processing 21

V. Conclusion 27

References 27

Part II Speech Perception

2 Articulatory and Acoustic Characteristics of Speech Sounds, Lucinda Wilder

I. Introduction 31

II. Production of Speech Sounds 33

III. General Acoustic Properties of Speech Sounds 36

IV. Articulation of Speech Sounds 43

V. Occurrence of Speech Sounds 46

VI. Vowel Phonemes of English 51

VII. Coarticulation 56

VIII. Consonant Phonemes of English 57

IX. Individual Differences 70

X. Distinctive Features 71

References 76

3 Acoustic Features in Speech Perception, Dominic W. Massaro

I. Introduction 77

II. Primary Recognition 78

III. Vowel Recognition 80

IV. Consonant Recognition 92

V. Conclusion 121

References 121

4 Preperceptual Images, Processing Time, and Perceptual Units in Speech Perception, Dominic W. Massaro

I. Introduction 125

II. Primary Recognition 125

III. Preperceptual Images and Processing Time 128

IV. Perceptual Units 137

V. Primary and Secondary Recognition 143

VI. Conclusion 149

References 149

5 Theories of Speech Perception, Kenneth R. Paap

I. Introduction 151

II. Summary of Information-Processing Analysis of Speech Perception 152

III. Descritions of the Theories 153

IV. The Role of Linguistic Structure in Model Building 175

V. Summary 199

References 200

Part III Reading

6 Visual Features, Preperceptual Storage, and Processing Time in Reading, Dominic W. Massaro and Joseph Schmuller

I. Introduction 207

II. Visual Features 210

III. Perceptual Storage 231

IV. Processing Time 234

References 237

7 Primary and Secondary Recognition in Reading, Dominic W. Massaro

I. Introduction 241

II. Utilization of Redundancy 242

III. Phonological Mediation 260

IV. Nonmediated Models of Reading 273

V. Mediated Models of Reading 278

References 286

8 Reading Eye Movements from an Information-Processing Point of View, Wayne Shebilske

I. Introduction: Significance of Reading Eye Movements 291

II. Characteristics of Reading Eye Movements 292

III. Oculomotor Control During Reading 301

IV. Summary and Conclusion 308

References 309

Part IV Psycholinguistics

9 Linguistic Theory and Information Processing, Kenneth B. Solberg

I. Introduction 315

II. Linguistics and Psychology 316

III. Theories of Grammar 318

IV. Semantics and Syntax 331

V. Models of the Language User 344

VI. Summary 352

References 353

10 Word and Phrase Recognition in Speech Processing, Arthur Freund

I. Introduction 357

II. Word Recognition and Context 359

III. Acoustic Cues of Grammatical Structure 367

IV. Click Localization 374

V. Summary 386

References 387

11 An Analysis of Some Psychological Studies of Grammar: The Role of Generated Abstract Memory, Joseph B. Hellige

I. Introduction 391

II. The Nature of Short-Term Memory (STM) 392

III. The Psychological Reality of Constituent Structure 394

IV. The Psychological Function of Certain Transformation Rules 413

V. Summary 421

References 422

Author Index 425

Subject Index 432

Seven Theses on Artificial Intelligence and Natural Language

Yorick Wilks
Department of Artificial Intelligence, University of Edinburgh

Institute for Semantic and Cognitive Studies, Fondazione Dalle Molle, Working Paper 17, 116 pages, 1975

Words and semantic primitives are not utimately of different types, although procedural benefits come from making the distinction. Plans stated in high level programming languages are most appropriately seen as texts, since they do not seem to be either real world procedures or procedures for handling natural language. As a general method of parsing natural language, expectation is radically defective unless it has some general and systematic capacity for attending to what it is reading. The important distinction between causes and reasons in the explanation of human behavior should have procedural and not only taxonomic reflction in an understanding system. Template to template (in preferential semantics) inference rules are: GOAL, CAUSE, IMPLIC. The CAUSE/GOAL distinction often retuces to no more than the temporal directionality of the rule. Real world knowledge can be represented quite usefully at quite local levels in an understanding and can function there as part of a general system of linguistic inference. The use of frames threatens to swamp us in large frame structures for relatively trivial matters and it is not clear how frames are related to lower level structures. We need multilevel representation. It is not at all obvious that a NL understanding system should be responsible or modeling general knowledge of the world.

SEMANTICS-DISCOURSE

On Natural Language Based Query Systems

Stanley R. Petrick
IBM J. Watson Research Center, Mathematical Sciences Department, Yorktown Heights, New York

IBM Research Report 5577, 13 August 1975

Some of the arguments which have been given both for and against the use of natural languages in questioning-answering (QA) systems are discussed. The following systems are considered in evaluating the current level of QA system development: LSNLIS, REL, SHRDLU, REQUEST There is a trade-off between syntactic and semantic complexity. A system with relatively simple syntactic capabilities must have complex semantic analysis procedures while a system, such as REQUEST, with sophisticated syntax can produce underlying syntactic structures which directly reflect meaning without the need for "creative" interpretation. A brief comparison between processing times in LSNLIS and REQUEST is given.

Question Answering in a Story Understanding System

Wendy Lehnert
Department of Computer Science, Yale University

Research Report 57, December 1975

This theory of question answering is based on the SAM (Script Applier Mechanism) mechanism. In the interpretative phase it takes a question in Conceptual Dependency form and categorizes it in terms of particular question types: 1) why, 2) how, 3) yes or no, 4) occurrence, 5) component. Each question type corresponds to a specific form of CD representation. In the response phase the memory is searched for the answer; this may involve nothing more than simple information retrieval or it may entail inferring the answer using general knowledge of the world. The system tries for completeness in answering; a yes/no question will be answered with yes or no *plus* an account of that answer. Work is being done on a Generation-Selection paradigm in which each question generates a number of feasible answers (the problem of memory representation) and a selection procedure chooses among them (pure QA). Selection rules are presented and discussed.

SEMANTICS-DISCOURSE: TEXT GRAMMAR

Beyond the Sentence, Between Linguistics and Logic

Janos S. Petofi
Universitaet Bielefeld

Hakan Ringbom, Ed., Style and Text, Sprakforlaget Skriptor AB, Stockholm, 377-390, 1975
ISBN 91-7282-095-0

The 'text-structure-world-structure theory' (TeSWeST) is an empirically motivated logic-oriented theory aiming at the grammatical description of a text as a complex sign (intensional semantic description) and the assignment of the possible extensional interpretations to the intensional-semantically described text structure (extensional-semantic theory). The intensional-semantic and extensional-semantic descriptions are such that they also contain the description of the pragmatic aspects. The grammatical component of the TeSWeST is a generative transformational text grammar operating with linearly not fixed canonic basic structures. The formation rule system of this grammar consists of a so-called communicative rule (R^C) expressed informally as: A communicative basis (TB^C) is a communicative predicate-complex; a communicator (C_1) communicates (COMM) to a (potential) communicator/interpreter (C_2) at a given time ($*Q^t$) in a given place ($*Q^l$) the message TB. The elements of the normed *implicit* representation of the text intension are definienda in the lexicon and the elements of the normed *explicit* representation of the text intension are definientes in the lexicon. The task of the extensional semantic component is to assign possible extensional semantic interpretations to the possible intensional semantic representations.

Linguistics and Artificial Intelligence

Petr Sgall
Centre of Numerical Mathematics, Charles University, Prague

Prague Bulletin of Mathematical Linguistics 24:5-33, 1975

(1) Winograd's approach to language has forced a reevaluation of the relationship between the theory of competence and the study of performance, pragmatics, etc., though it is still wise to acknowledge the descritption of language as a relatively independent enterprise. (2) The use of linguistics descriptions in AI provides a test for linguistic theories. (3) Winograd's "imperative form" of representing knowledge and semantics is more effective than one based entirely on deductive logic. Considerations of topic and focus, theme and rheme, functional sentence perspective, etc. are similarly "imperative" in their emphasis on the different "uses" which various items of information in a communication have. (4) Winograd's work, in his way of introducing new definitions into the semantic component, suggests that in man-machine communication the burden of learning the other participant's language may be shiftable from man to the computer. (5) The significance of linguistics for AI is connected with making programming languages more and more like NL's. (6) The study of tense and time and the study of negation allow us to see how linguistic meaning and non-linguistic content can be distinguished.

Formal Philosophy: Selected Papers of Richard Montague

Richmon H. Thomason, ed.
University of Pittsburgh

Yale University Press, New Haven, Connecticut, 1974. ISBN 0-300-01527-5
HC: $12.50

Contents

Introduction, By Richard Thomason 1

1. Logical Necessity, Physical Necessity, Ethics, and Quantifiers 71

2. 'That' (with Donald Kalish) 84

3. Pragmatics 95

4. Pragmatics and Intensional Logic 119

5. On the Nature of Certain Philosophical Entities 148

6. English as a Formal Language 188

7. Universal Grammar 222

8. The proper Treatment of Quantification in Ordinary English 247

9. A Paradox Regained (with David Kaplan) 271

10. Syntactical Treatments of Modality, with Corollaries on
Reflexion Principles and Finite Axiomatizability 286

11. Deterministic Theories 303

Works of Richard Montague, a Bibliography 361

Index 365

More on In-Core Sort-Search Methods in PL/1 for Lexical Data

Richard K. Brewer
Eastern Michigan University, Ypsilanti

SIGLASH Newsletter 8, No. 1:9-12, February 1975

Four procedures based on binary tree structures are presented along with PL/1 code. 1) Symmetric or inorder traversal (Knuth) exploits the powerful syntax of PL/1 in the use of modular code and recursion. 2) Explicit stack (Knuth). 3) "Threading" a tree (Perlis & Thornton) permits traversal without giving up space for either an implicit or an explicit stack. 4) This technique requires an estimate of outer bound for the size of the list; by allocation of a fixed binary array at the outset the tree can be constructed by basing fresh nodes on progressively higher elements in the array.

COMPUTATION: PROGRAMMING

SNOBOL4 Applications in Natural Language Research

James L. Wyatt
Florida State University, Tallahassee

SIGLASH Newsletter 8, No. 2-3:12-19, April-June 1975

Topics discussed: 1) Pattern matching basics 2) the equals mark to create or alter character strings, 3) statement format, 4) INPUT and OUTPUT, 5) simple programs. A simple program to count word frequencies in a text is described and discussed and 12 examples of output from student programs are given.

Revised Report on the Algorithmic Language ALGOL 68

A. van Wijngaarden, B. J. Mailloux, J. E. L. Peck, C. H. A. Koster, M. Sintzoff, C. H. Lindsey, L. G. L. T. Meertens, and R. G. Fisker, Eds.
International Federation for Information Processing

Acta Informatica 5:1--236, 1975

Contents

Acknowledgements 6

0. Introduction 8

PART I: Preliminary Considerations

1. Language and metalanguage 17

2. The computer and the program 35

PART II: Fundamental Constructions

3. Clauses 53

4. Declarations, declarers and indicators, 66 . . . .

5. Units 77

PART III: Context Dependence

6. Coercion 91

7. Modes and nests 98

PART IV: Elaboration-independent constructions

8. Denotations 108

9. Tokens and symbols. 113

PART V: Environment and Examples

10. Standard Environment 124

11. Examples 209

12. Glossaries 218

Interacive Graphics on Intelligent Terminals in a Time-Sharing Environment

W. K. Giloi, and S. Savitt
University of Minnesota, Minneapolis, 55455

J. Eucaruacao
Fachbereich Informatits, EG Graphiscke Datenverarbeitung, Technische Hochschule Darmstadt, D-61, Germany

Acta Informatics 5:257-271, 1975

Interactive graphics in a time-sharing environment should be organized in such a way that the user's activities are locally processed in order to avoid unacceptably long response times. On the other hand, the host computer must be kept informed about the user's actions and, ccnversely, the display file in the terminal has to be updated whenever the execution of the application program causes a change in the visual representation. In order to avoid the transmission of redundancy, the display file is decomposed into two intersecting parts such that the part in the host computer and the other in the terminal contains only the locally required information. The necessary communication between both parts is maintained by an information module generated on the base of a low-low-level intermediate language (L^4) and exchanged between computer and terminal. This leads to the notion of an abstract terminal whose "machine language" is L^4, facilitating the implementation and portability of graphic programming systems.

A Multilevel Modeling Structure for Computer Generated Graphical Symbol Design

Chuan Lee
McDonnel-Douglas Astronautics Company, Huntington Beach, California

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 387-406

A multilevel data structure has been developed to allow its use to construct graphical symbols on a automated drafting machine. The implemented logic allows a user to define several sets of alphanumerical characters and graphical symbols in terms of three levels of tables for creating and retrieving the needed line segment strokes. It is believed that the same logic could be applied to provide a solution to the computer output of Chinese characters. To call any desired Chinese character as output, it is simply a matter of calling the character's legal designator within a particular library.

Interactive Graphical Data Analysis with Implementation of Chinese Characters

Karen K. Yuen
Department of Biomathematics, University of California, Los Angeles

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 373-377

Treating Chinese characters as pictures we can display them on the screen without difficulty. Subroutines are written to generate each word or part of the word. On the input side, one can use a light-pen on the appropriate part of the screen to choose which part of the program one wants to go to, e.g., go to next frame, delete this point, etc. A push of the designated function key switch will also convey information. Chernoff proposed the mapping of multidimensional variables onto features of faces. Mouth, nose, eyes, facial contour, etc., are modified in form and size to represent each vector of measurements. A graphics program was written using this idea and extending it to provide the histogram comparison and interactive classification of individual cases.

DOCUMENTATION

Two Major Flaws in the CODASYL DDL 1973 and Proposed Corrections

G. M. Nijssen
Control Data Europe, 46, Avenue des Arts, B-1040, Brussels, Belgium

Information Systems 1:115-132, 1975

In a schema written in CODASYL DDL 1973 it is syntactically correct to describe an entity either by declaring a data item in the record of the entry or by declaring a CODASYL set type in which the record, describing the entity, is a member record. This is a major flaw because extension of a database or integration of two existing databases will then lead to either reprogramming or inconsistency, or both. The flaw can be corrected by requiring that all attributes are represented as a data item in the (logical) schema. In the CODASYL DDL 1973, there are five places to optionally declare a record identifier and four of these five places are not in the record but in the CODASYL set type. Declaring record identifiers therefore results in fairly complex and non-orthogonal declarations. This could be simplified by abondoning these five places and by introducing a record identifier clause in the record type entry. For integrity reasons it is necessary to require that at least one record identifier is declared in every record type entry. The previous two corrections will make is possible to design a CODASYL set selection clause, still providing the same functional capabilities. The corrected DDL is functionally equivalent, yet offers more data independence, is simpler and more orthogonal. Examples are given.

Allocating Storage in Hierarchical Data Bases Using Traces

P. A. Bernstein, and D. C. Tsichritzis
Department of Computer Science, University of Toronto, Canada M5S 1A7 *Information Systems 1:133-140*

Traces can be used to allocate a hierarchical (i.e. tree structured) data base. A trace is an n tuple of indices, $[x(1),...,x(n)]$, which describes the unique path from the root of the tree to the node being addressed. That is, one takes the $x(1)$-th branch from the root, followed by the $x(2)$-th branch from the next node, etc. until the path is completed. The last node on the path is the one being addressed. Given a set of traces that represent a set of nodes in a tree, the problem is to allocate them efficiently on a file. We approach the problem by finding ways of mapping n-tuples (i.e. traces) onto natural numbers (i.e. file indices). An allocation scheme is proposed which uses a 1:1, onto trace-to-address map and is designed to adapt to a changing distribution of nodes within the tree.

DOCUMENTATION

A General Model of Distribution of Objects in Information Retrieval Systems

Miroslaw Dabrowski
Institute for Organization, Management and Control Sciences of the Polish Academy of Sciences, 00-818 Warsaw, 55 KRN Street

Informations Systems 1:147-151, 1975

Hsiao and Harary have given a generalized file structure and shown that the three most commonly used types of file structure: inverted, index-sequential, and multilist files, are all special cases of a generalized file structure. Other methods allow the creation of buckets of objects by means of the combination of descriptors. A general model of informatiion structure is proposed in which the concept of distribution is introduced and different kinds of distribution of a set of objects are considered: 1) Inverted, 2) Which corresponds to those methods of information retrieval in which every bucket description is a conjunction of k non-negated descriptors ($1<k>n$), 3) Canonical, in which a redundancy equals to zero, 4) Which corresponds to the index sequential method, 5) Modified Canonical. Well-known methods of information retrieval are special cases of the general model.

The Definition of Concepts

Fred W. Riggs
University of Hawaii

Occasional Paper No. 6, International Studies Association, University Center for International Studies, University of Pittsburgh, 39-76, 1975

Definitions should exploit the distinction between *definiendum* and *definiens.* The definiens designates a concept *in context* while the definiendum defines the concept *without dependence* on context. The definiendum of one concept X_1 may appear as a term in the definiens of another concept X_2, thus permitting recursive construction of definitions. A machine based archive of social science concepts is proposed in which the core element of each entry will be a definition copied from relevant scholarly writings. Each entry will also contain: a) a possible translation of the definition into English and into formal precise language b) specification of related terms and relevant theoretical context, c) documentation of the source of the difinition, d) name of the contributor, e) anything else which seems appropriate. Terms found in the new Thesaurus of the American Political Science Association will be used from the beginning with later additions as needed. The project will be administered through the Information Utilization Laboratory at the University of Pittsburgh.

DOCUMENTATION: CLASSIFICATION

Word Segmentation by Letter Successor Varieties

Margaret A. Hafer, and Stephen F. Weiss
Department of Computer Science, University of North Carolina, Chapel Hill

Information Storage and Retrieval 10:371-385, 1974

Within a word, the *i*th letter is to some degree dependent on the *i*-1 letters that precede it. Taking advantage of this, successor and predecessor letter variety counts are used to indicate where words should be divided. Four segmentation strategies are used: a) Cutoff, b) Peak and plateau, c) Complete word, d) Entropy. Experiments have been run testing the machine implementation of 'various combinations of of these strategies. A technique involving whole word word segmentation techniques and cutoff points on successor and predecessor counts was chosen for use in information retrieval experiments. The information retrieval results obtained are virtually identical to those obtained with the more manually oriented forms of stemming.

Application of Minimum Spanning Trees to Information Storage

R. C. T. Lee, and C. L. Chang
Heuristics Laboratory, Division of Computer Research and Technology, National Institutes of Health, Department of Health, Education, and Welfare, Bethesda, Maryland

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 1245-1256

The use of minimum spanning trees allows efficient storage of information but permitting information which is redundant across similar items to be stored at only one point in the system. Let the distance between two points be a function of the number of changes of values between these two points. A minimum spanning tree contains all the points and is such that its total sum of distances is a minimum among all such possible trees. Examples discussed: classification of animals based on amino acid sequences, voting records of countries in the UN. Applications to Chinese characters are discussed.

DOCUMENTATION: RETRIEVAL

An Efficient Retrieval Method for a Hierarchical Fact-Retrieval System and Its Evaluation

Fujio Nishida, and Shinobu Takamatsu
Faculty of Engineering, University of Osaka Prefecture, Sakai-shi, 591 Japan

Systems-Computers-Control 6, No. 1:52-60, 1975

A fact-retrieval system consisting of a data base, axiom set and inference system can deduce answers to questions by combining a certain number of facts. A predicate symbol followed by a series of constant terms or variable terms and its complement is a literal. A literal without variable terms is a ground literal. A clause is a logical sum of literals. A hierarchical fact-retrieval system is studied by viewing data as a set of clauses, each of which consists of a ground literal. Concrete fact-retrieval is assumed not to contain function symbols by limiting the objective. Under such condition efficient procedures are given for: 1) retrieving constant terms by specifying predicate symbols and part of the constant terms and, 2) retrieving predicate symbols and the remaining constant terms by specifying part of the constant terms. Futhermore, the number of comparisons, which is the criterion for the retreival time in each procedure, is examined analytically and experimentally.

A Linguistic Approach to Information Retrieval--I

Petr Sgall, and Eva Hajicova
Centre of Numerical Mathematics, Charles University, Prague

Information Storage and Retrieval 10:411-417, 1974

The system should consist of: 1) the brain, which organizes the whole repertory of data, 2) the analysis, which takes, as its input, text and questions in NL mixed with formal notations, and presents, as output, disambiguated translations in a language the brain can directly use, and 3) the synthesis, translating the answers delivered by the brain into the appropriate NL. The representation of the sentence on the tectogrammatical level (the artifical language) has the form of a dependency graph, with the verb as its root. The participants of the verb are ordered according to: 1) an inherent order of participants determined by the language system, 2) topicalization and communicative dynamism, 3) rules of grammar. The semantic representation is marked as to which elements are contextually bound and which are not. The verb always stands between its contextually bound and contextually non-bound participants.

A Linguistic Approach to Information Retrieval--II

Petr Sgall, Jarmila. Panevova, and Svatava Machova
Centre of Numerical Mathematics, Charles University, Prague

Information Processing and Management 11:147-153, 1975

Synthesis of Czech in MT. The generative component is a context-free phrase structure and uses modifying, substitutional and selectional rules. The right-hand side of the rules cannot contain more than two non-terminal symbols. There are recursive rules. The order of application is determined by the form of the rule itself (by a selection of non-terminal symbols). The rules are not distinguished as obligatory and optional ones. The translation of the semantic representation to the graphemic level proceeds in two steps. First, the semantic representation is translated into a surface syntactic level, and this is then (secondly) translated into a morphemic representation. The sequence of computer programs is based on the formal pattern of pushdown transducers.

Some Computer Functions for Machine-Aided Translation

J. Mathias
GETA Group, U.S.A.

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 589-592

The experimental computer configurations use alphanumeric input by keyboard (and simulation of graphic input). Both systems depend on retrieval from a glossary file of Chinese characters with English meanings and a cross reference file of characters containing transliterations. The first set of computer functions put together is based on the prime method of input by alphanumeric keyboard and graphic output for displaying Chinese characters. The most important of these is the segmenting function, which is capable of searching the glossary file for meanings for each individual character and for sets of contiguous characters in the query string. The system has two indexing functions: partial telecode indexing and Pin Yin indexing. The second set of computer aid functions are based on input by Pin Yin romanization. The operator can type in the Pin Yin spelling (without tone) for a query of up to seven characters in length and the system searches the glossary file for all terms that fit the Pin Yin. This method is effective only for compound queries.

TRANSLATION

Mechanical Translation Between English and Japanese

Shigeharu Sugita
Department of Information Science, Kyoto University, Yoshida Sakyo-Ku, Japan

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 555-572

Phrase structure grammar is supposed for both English and Japanese, and a syntax to syntax translation by ordered context-free grammar is adopted which concerns word order exchange and insertion and deletion of auxiliary particles, and does not concern the semantical aspect of translation words. The main part of the algorithm, syntax analysis and synthesis, is carried out by using context-free type rewriting rules. These rules are classified into several hierarchies and there exist priorities in each hierarchy when rules are used. Program and linguistic data are separated as much as possible, so that a change of grammar does not affect the main program. Therefore the program size becomes very small: about 1500 statements in assembler language. The word dictionary contains 8000 head words and the grammar table contains about 900 rewriting rules. This translation system can in principle accept any complex structure, and about 70% of sentences from scientific or technological papers are analyzed correctly from the syntactical point of view. Results of English into Japanese translation can be spoken through speaker by voice synthesizer.

Tower of Babel: On the Definition and Analysis of Concepts in the Social Sciences

Giovanni Sartori
University of Florence

Fred W. Riggs
University of Hawaii

Henry Teune
University of Pennsylvania

Occasional Paper No. 6, International Studies Association, University Center for International Studies, University of Pittsburgh, Pennsylvania, 1975

Contents

Preface i

Introduction: Toward the Mastery of Obstacles to Conceptual Clarity

1. Obstacles 1

2. A Broad Strategy 2

1) conceptualization; 2) hierarchies; 3) the "paradigm" level

3. Summary and Conclusion 5

4. Next Steps: Plans and Prospects 5

Chapter 1: The Tower of Babel, Giovanni Sartori

Introduction 7

1) the loss of etymological anchorage; 2) the loss of historical anchorage; 3) the loss of mainstream discourse; 4) novitism; 5) the freezing of language; 6) the cards and the game; 7) counteracting chaos; 8) the paradigm-model juncture

2. Concepts, Words, Phenomena 12

1) which is the start? 2) language and thinking; 3) the impact of words; 4) the neo-Baconian start

3. Conceptual Analysis 15

1) terms as concepts and terms in sentences; 2) logical systematizing; 3) analysis by classification; 4) vertical classification; 5) rules of transformation along a ladder of abstraction; 6) an illustration; 7) concepts as data containers; 8) theory poor and data cheated

4. Measurements and Quantification 21

1) the broad meaning; 2) the narrow meaning; 3) kinds and degrees; 4) objections; 5) pros and cons; 6) quantities of what?

5. Applied Logic 25

1) science and logic; 2) strategy of allociation; 3) truth-operators and quantifiers; 4) theoretical terms and observational terms; 5) object concepts and property concepts; 6) entities should not be multiplied; 7) the allocation of dispositional terms; 8) simple definitions; 9) varieties of definitions; 10) operational definitions; 11) minimal definitions; 12) arbitrariness in defining

6. Hindthoughts 36

1) Heraclitus versus Descartes; 2) the missing bridge

Chapter Two: The Definition of Concepts, Fred W. Riggs

1. The Definition of Definition 39

1) words and dictionaries; 2) dictionary definitions; 3) the definition of definitions; 4) lexical definitions; 5) lexicographical propositions; 6) analytic definitions; 7) defining propositions; 8) statements; 9) nominal definitions, 10) prescriptive definitions; 11) stipulative definitions; 12) summary of definitions; 13) so-called definitions

2. The Definition of Concepts 46

1) the ambiguity of nominal definitions; 2) the dictionary definition of concepts; 3) definitional functions; 4) terminological agreement; 5) the origin of concepts; 6) explicative definitions; 7) strategies for concept formation; 8) the explicative strategy; 9) the constructive strategy; 10) evaluating the alternative; 11) natural science precedents; 12) summary: complementary approaches; 13) reconstruction vs. explication; 14) a choice-concept puzzle?

3. The Transformation of Concepts and Definitions 54

1) symbols and the symbolized; 2) properties of definitions vs properties of concepts? 3) the definiendum as concept designator; 4) the definiens as concept connoter; 5) some implications; 6) transformation of definitions; 7) toward significance and concreteness; 8)changes on the ladder; 9) feedback from relevance; 10) terminological confusion from unrecognized concept changes

4. The Clarity of Definitions: Parsimony 63

1) a necessary but not sufficient condition; 2) on the optimality of definitions; 3) parsimony in definitions; 4) simple redundancy; 5) logical redundancy; 6) empirical redundancy; 7) indicators and redundancies; 8) detectors and measures

5. The Adequacy of Definitions 70

1) an illustration: red; 2) concepts as defining criteria; 3) properties and dimensions; 4) intrinsic and extrinsic properties; 5) adequacy of definitions: the status of concepts

6. A Concept Inventory 73

1) an open archive; 2) sources of concept definitions; 3) priorities for concept selection, 4) inventory financing; 5) retrieval technology; 6) inventory utilization; 7) long-term strategies

Chapter Three: On The Analysis of Concepts, Henry Teune

1. Introduction 77

2. Science and Philosophy of Science 78

1) divergence between philosophy and science; 2) the role of philosophy of science; 3) inadequacy of philosophy of science for social science; 4) some reasons for the inadequacy

3. Concepts and Definitions 79

1) meanings of concepts; 2) value concepts; 3) logical concepts; 4) empirical concepts; 5) definitions and concepts

4. Evaluating Concepts 80

1) criteria for evaluating concepts. 2) empirical precision; 3) theoretical importance; 4) relationship between empirical precision and theoretical relevance

5. Composition Rules 82

1) simple and complex definitions; 2) the structure of composition rules; 3) the theoretical importance of composition rules; 4) empirical limits on the use of composition rules

6. Object Concepts 83

1) the nature of object concepts; 2) importance of object concepts in the social sciences; 3) the cutting edge of object definitions; 4) object definitions and near tautologies. 5) relationships between object and property concepts

7. Types of Object Concepts: Aggregated Systems and Levels 86

1) various meanings of levels; 2) object concepts and generality; 3) classes of objects; aggregates; 4) systems and classes of systems

8. Differentiating Types of Property Concepts 88

1) according to function; 2) according to theory; 3) according to the structure of definition; 4) according to the characteristics of the objects; 5) according to properties; 6) according to the nature of the composition rule; 7) some miscellaneous types of concepts

9. Measurement 91

1) measurement contrasted with definitions; 2) measurement languages,

operations, and statements; 3) two types of measurement; 4) two traditions of measurement; 5) indicators

10. A Concluding Statement 93

Conclusion, Fred W. Riggs 95

Footnotes 99

*This chapter has been abstracted in the Documentation section of this fiche.

SOCIAL-BEHAVIORAL SCIENCE: ANTHROPOLOGY

On Binary Categories and Primary Symbols: Some Rotinese Perspectives

James J. Fox
Harvard University

Roy Willis, ed., The Interpretation of Symbolism, Halsted Press Division John Wiley & Sons., 1975, 99-132. ISBN 0-470-94920-1
HC: $17.75

Rotinese chants are ordered by extensive parallelism. Each opposition in the Rotinese ritual language is a *dyadic* set. Some terms may appear in more than one dyadic set. Tracing relations among these semantic elements involves chains and cycles along the edges of a symmetric graph. The analysis of 5000 lines of verse has yielded a dictionary of the ritual language consisting of 1000+ dyadic sets over 1400 entries. By concentrating on elements included in 5 or more dyads (the frequency cutoff is somewhat arbitrary) and with at least 2 links in common a core set of 21 entries is identified. The core includes directional coordinates, words for earth', 'water', 'rock , and 'tree" terms for plants and plant-parts, body parts, and a peculiar collection of verbs of position involving ideas of balance, border, ascent, and descent. This core has been investigated using cluster analysis programs.

On the Complexity of Causal Models

B. R. Gaines
Man-Machines Systems Laboratory, Department of Electrical Engineering Science, University of Essex, Colchester, England

IEEE Transactions of Systems, Man, and Cybernetics 6:56-59, January 1976

The principle of causality is fundamental to human thinking. It has been observed experimentally that causal thinking leads to complex hypothesis formation by human subjects attempting to solve comparatively simple problems involving acausal randomly generated. The assumption of causality in modeling acausal systems leads to meaningless models that cannot reflect any stochastic structure present. This correspondence provides an automata-theoretic explanation of this phenomenon by analyzing the performance of an optimal modeler observing the behavior of a system and forming a minimal-state model of it.

SOCIAL-BEHAVIORAL SCIENCE: PSYCHOLOGY

Computer-Determined Readability Profiles

David M. Locke
Illinois Institute of Technology, Chicago

Alan K. Stewart
Illinois Institute of Technology Research Institute, Chicago

SIGLASH Newsletter 8, No. 4:8-12, October 1975

The readability measure is based on sentence length in words and word length in alphanumeric characters. The program measures readability on paragraphs and paragraph blocks and proceeds paragraph by paragraph, cumulating data on analyst assigned blocks and over the entire passage. Paragraphs can be skipped if necessary. An example of the program's operation is given.

Computer Simulation of a Language Acquistion System: A First Report

John R. Anderson
Human Performance Center, University of Michigan

Robert L. Solso, Ed., Information Processing and Cognition: The Loyola Symposium, John Wiley & Sons, 295-349, 1975

LAS (Language Acquistion System) is an interactive program which accepts as input lists of words, which it treats as sentences, and scene descriptions encoded in a variant of the HAM propositional language. It utilizes an augmented transition network grammar (Woods) for both parsing and generation. The SPEAK program starts with a HAM network of propositions tagged as to-be-spoken and a topic of the sentence. SPEAK uses a depth-first strategy to find a path in the parsing network which can be used to generate the requested sentence. In the UNDERSTAND program it is necessary, when one path through the network fails, to consider the possibility that the failure may be in a parsing of a subnetwork called on that path and to go back into the network to attempt a different parsing. In acquistion, BRAKET is an algorithm for taking a sentence of in arbitrary language and a HAM conceptual structure and producing a bracketing of the sentence that indicates its surface structure. This surface structure prescribes the hierarchy of networks required to parse the sentence. After BRACKET is complete, SPEAKTEST is called to test whether its grammar is capable of generating a sentence and, if it is not, appropiately to modify the grammar so that it can. A list is kept of all networks created by SPEAKTEST; GENERALIZE is then called to determine which networks are identical. At present LAS operates in a highly restricted semantic domain, but a more challenging domain is planned for future work.

SOCIAL-BEHAVIORAL SCIENCE; PSYCHOLOGY: LEARNING

A Mathematical Theory of Learning Transformational Grammar

Henry Hamburger, and Kenneth Wexler
School of Social Services, University of California, Irvine

Journal of Mathematical Psychology 12:137-177, 1975

The language-learning theory has four interrelated aspects: (a) the class, G, of possible grammars, (b) the kind, I, of information made available to the learner, (c) the language-learning procedure, P, and (d) the criterion, C, of success. The theory must deduce a guarantee that the correct member, g, of G can be discovered by applying procedure P to information of type I about g. The sense in which g is "discovered" must be made precise by a formal criterion C, and the proof must hold for arbitrary g in G. The system presented converges; that is, the learning process learns the language to a formal criterion. The claim that the theory is reasonable rests on the plausibility of G, I, P, and C. G = the class of all TG's. Each element of I is a sentence coupled with its underlying structure. The procedure P doesn't require explicit memory for past data and must acquire the ability to map each phrase-marker (akin to meaning) to the appropriate surface structure.

Open-Ended Application of Information Processing Techniques in the Humantities

Richard S. Thill, and Jerry L. Ray
University of Nebraska, Omaha

SIGLASH Newletter 8, No. 2-3:7-11, April-June 1975

A system to store, manipulate, and retrieve linguistically structured material in interactive usage has been applied to the poetic corpus of Heine. The system can: 1) present word occurrence frequently lists alphabetically and by frequency, 2) is provided with a table of contents about material stored on the tape, 3) has a word occurrence index with volume (in the printed edition) and page references and frequency of occurrence, 4) provide key word occurrence lists--the researcher lists words or stems of interest and receives a list of contexts with volume, page, and line numbers, 5) provide a chronological histogram--researcher specifies term or terms for examination and is given a graphic display of variation in word class frequency through time. Applications to CAI, Bibliography, and Reference materials are being developed.

HUMANITIES

The Computer and Creativity

Jay A. Leavitt
University of Minnesota, Minneapolis

Allen R. Hanson
Hampshire College, Amherst, Massachusetts

SIGLASH Newsletter, 8, No. 4:4-7, October 1975

Though the computer itself cannot be considered creative, the artist can use the computer in 2 modes: 1) as an active device contributing through simulation, to the final artistic product, and 2) passively, like a tool to be manipulated by the user. Once we think we understand the structure of a class of objects under consideration, that is, how the parts making up the whole are interrelated, we can actively use this information and the random element to "create" objects similar to those being studied.

Wandering in mist
Reaching out to soft sunlight
Blue-scaled dragons pause.

Sasanian Pahlavi Inscriptions: A Concordance

J. A. Moyne
Department of Computer Science, Queens College, Department of Linguistics, Graducate Center of the City University of New York

Computers and the Humanities 8, 27-39, 1974

The collected volume contains all the known inscriptions in Pahlavi or Middle Persian of the Sasanian era (A.D. 226 to A.D. 651). The collection of 21 inscriptions was transliterated and punched on 718 IBM cards used as input for the concordance program, which produced the following output material: 1) a complete listing of the texts on the inscriptions, each separated under its heading, 2) Two alpabetized concordances, generated from all inscriptions which are treated as one text, one for Pahlavi and one for Aramaic words, 3) Two indices or listings (one for Pahlavi, one for Aramaic) produced for the words in the input corpus with line references following each word. This general-purpose concordance and text-processing computer program can be used for similar productions for any text in any language for which a suitable transliteration convention is available. Examples.

HUMANITIES: ANALYSIS

Pericles and the Question of Structural Unity

T. R. Waldo
University of Florida, Gainsville

SIGLASH Newsletter 8, Nos. 2-3:22-27, April-June 1975

The analysis of *Pericles* by computer establishes the organic function of ideas clustered around themes of chance (external forces operating on man), with 927 occurrences, and choice (humanly controllable forces), with 979 occurrences. An occurrence table indicates, quantity of occurrence, distribution of themes through scenes, and proportion of occurrence of a theme--defined in terms of the percentage of words in scene with a certain classification symbol (a theme indicator) divided by the percentage of words of that class in the play as a whole. The use of the thematic profile must be supplemented with conventional techniques of literary analysis.

Mathematical Theory of Free Rhythm

Evzen Kindler
Department of Mathematical Informatics, Faculty of Mathematics and Physics, Charles University, Prague, Czechoslovakia

Kybernetika 11:221-233, 1975

Free rhythm is that used in ancient Greek and Latin prose and in the music of the early Middle Ages in Europe. An elementary arsis(A_E) is an elementary time value plus an upbeat; an elementary thesis(T_E) is one or two elementary time values plus a downbeat. An elementary rhythm(R_E) is: $A_E T_E$. A simple rhythm consists of a simple arsis (not the same as an elementary arsis) followed by a simple thesis. Rhythms may be recursively composed of other rhythms to form a phrase(K). A phrase is a result of the highest rhythmical synthesis of the elementary time values while the ordering of phrases is no more an affair of rhythm but of stylistic form. A grammar of free rhythm is an ordered 4-tuplet: G = (V, V_1, K, S), where V is a set of 6 terminal symbols, V_1 is a set of 13 auxiliary symbols, K is a member of V_1 and is the initial symbol, and S is a set of 21 production rules. A modified grammar(G*) has 5 terminal symbols, 6 auxiliary symbols, and 13 production rules, but offers an equivalent rhythmic synthesis. An algorithm is given to complete texts generated by G* and transform them into texts generated by G.

roving Musical Theorems I: The Middleground of Heiniich Schenker's Theory of Tonality

Michael Kassler
Basser Department of Computer Science, School of Physics, The University of Sydney

Technical-Report No. 103, August 1975

Schenker's theory of tonality asserts that every composition that is an instance of tonality can be derived from one of three *Ursaetze* (background) by the successive application of a small number of rules of inference called 'prolongation techniques' (middleground). Harmonic, rhythmic, melodic, dynamic, etc. details belong to the foreground stage of derivation. The middleground rules for the major mode involve two formalized languages, S_1, which has 11 inference rules (corresponding to Schenker's prolongation techniques), and S_2, with 5 inference rules. S_1 governs *dilynear* compositions (a *lyne* being one musical voice) and has three primitive axioms, while S_2 governs *trilynear* compositions and contains an axiom structure which relates the systems S_1 and S_2. The decision procedure provides one proof for any given theorem (i.e composition) rather than all possible proofs. At the middle level it seems that alternative minimal proofs in no way demonstrate a semantically important ambiguity, though the situation is likely to change significantly with the explication of foreground structures.

Rhythmic Variation in the Use of Time References

Jan Svartvik
University of Lund

Haken Ringbom, Ed., Style and Text, Sprakforlaget Skriptor AB, Stockholm, 416-432, 1975
ISBN 91-7282-095-0

A study of time references in James D. Watson's *The Double Helix* reveals that they occur rhythmically throughout the book, which has 29 chapters. If one graphs the percentage of time references per chapter against chapter numbers, the resultant curve has five peaks (Chapters 1, 8, 15, 22, and 29) with six chapters between successive peaks. There are four troughs (Chapters 2, 10, 17, and 25). The first and last gaps between the troughs cover seven chapters each while the middle gap covers six chapters. A similar rhythmic variation in time reference usage appears in the paragraph structure within individual chapters.

INSTRUCTION

Computer-Assisted Tutorial in College Mathematics

J. L. Caldwell
Department of Mathematics, University of Wisconsin-River Falls, 54022

Douglas Polley
Department of Mathematics, University of Minnesota, Minneapolis, 55455

Computers and People 24; No. 12:22-23, December 1975

The programs deal with quadratic equations and equations of straight lines. The student, seated at a teletype, receives a problem with ramdomly chosen parameters. The student then solves the problem and enters his solution. If his solution agrees with the computer's solution, then another problem is presented. If the student's solution is incorrect the computer then works the problem incorrectly in several different ways, each time checking its answer against the student's. If a match is obtained, then the computer suggest that the student has made a certain error and asks him to try again. If none of the incorrect answers match the student's answer, then the computer will check the student's work step by step. If a student requires detailed solutions to several problems, he is asked to contact the insttuctor.

Computer Applicability in a Programmed Instruction System for Chinese/ Japanese Characters

Benjamin K. T'sou, and Yat-Shing Cheung
Linguistics Department, University of California, San Diego

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 641-650

The set of video equipment consists of a video recorder with T.V. monitor, a specially modified camera and an electronic interphase device called the Genloc System. The camera is installed under the monitor in a cabinet and aims at a piece of translucent glass fitted onto a protruding panel. The student places a piece of paper on top of the glass and practices writing by copying the model on the screen. The system has three basic components: 1) Display, 2) Input of instructional material, which is organized into three distinct phases, and 3) Student input. When the student turns the switch for the Concurrent Display Mode, the character produced by him will be superimposed onto the instructor produced character and concurrently displayed on the screen. In the simulated computer aided system, the Concurrent Display Mode will be monitored not only by the student, but also by a research assistant through a remote monitoring unit. The research assistant will do what a computer would be expected to do.

INSTRUCTION

Teaching and Learning Chinese in a Computer Environment

Susan Poh
The MITRE Corporation, Westgate Research Park, McLean, Virginia

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 617-639

People in Western civilization encounter great difficulty when attempting to learn the Chinese language. The major difficulties seem to be due to the non-alphabetic, non-phonetic and pictorially-structured characters. The system described applies computers and computer graphics to teach the Chinese ideogram in a systematic, simple and effective manner. A set of LOGO procedures are defined which draw simple strokes and then these procedures are used as primitive commands to construct ideograms on a CRT, one stroke at a time. Five demonstration courseware lessons are discussed to illustrate the teaching methodology.

Computer Aided Instruction in Chinese Characters, II-Implementation

Shang-Chun Chen, and Henry Y. H. Chuang
Computer Systems Laboratory, Washington University, St. Louis, Missouri

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academic Sinica, 605-616

In character coding for display, the system uses a method based on "piecewise linear approximation" The procedure for stroke recognition has three steps: 1) Transform the stroke code into a "direction sequence". 2) Transform the direction sequence into "major direction sequence". 3) Identify the stroke type by processing the major direction sequence with an automaton. The structure of a character can be unambiguously described as a context-free grammar with strokes as the terminal elements of the grammar. The nonterminals include 'units' and 'components'. A unit consists of a group of closely related strokes and components consist of two units or of two smaller components. Based on this structure, the computer builds a structured base for error analysis. This base consists of a set of relation matrices, one for each production of its grammar. Error detection proceeds by comparing entries in the relation matrices of the model character with those of the character written by the student.

Computer Aided Instruction in Chinese Characters, I--The System

Chuang Henry Y. H., and Shang-Chun Chen
Computer Systems Laboratory, Washington University, St. Louis, Missouri

S. Gould, Ed., Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Academia Sinica, 599 603

The system functions completed so far include: 1) *Teach* the pronounciations, meanings, and the written form of characters, as well as how to write them, 2) *Guide* the writing, stroke by stroke, with or without information regarding the relative size and position of the strokes, 3) *Monitor* the student's writing to assist the student in correcting mistakes by himself, and 4) *Analyze* the student's writing and then indicate errors and how to improve.

Brains, Robots, and the Evolution of Language

Michael Arbib
Department of Computer and Information Science, University of Massachusetts, Amherst

Technical Report 74C-2, 40 pp.

For interacting with the world animals and robots need a spatical framework, the ability to segment the world into chunks, and long-term and short-term models of the world. The ability to see what objects are in the world and to interact appropriately with the world is shared by all animals. Language is not a magic separate device to be explained by formal grammars, but is rather an ability which evolved naturally out of our ability to perceive the world. The child's acquisition of language builds on his cognitive capacities and uses a strategy in which he reconstructs sentences according to his own rules and proceeds by fine-tuning of linguistic structures. A distributed information processing machine (DIPM) is suggested as a theoretical model in which the logic functions of the machine are distributed and not centralized. Consequently no single center exists in which damage could lead to the breakdown of the whole machine. The cybernetic study of language needs a theoretical language which encompasses AI work and brain theory.

BRAIN THEORY

Artificial Intelligence and Brain theory: Unities and Diversities

Michael A. Arbib
Computer and Information Science Department, Center for Systems Neuroscience, University of Massachusetts, Amherst

Technical Report 75C-8, 62 pages, September 1975

In the control of *movement*, AI offers insight into overall *planning* of behavior; while control theory enables BT (Brain Theory) to model *feedback* and *feedforward* adjustments by the spinal cord brainstem and cerebellum. A *schema* is an internal representation of an 'object' and comprises input-matching routines, action routines, and competition and cooperation routines. The internal representation of the world is then given by a 'collage' of tuned and activated schemas. A number of studies in AI and BT are discussed which offer hope of a unified theory of *competition and cooperation within a single subsystem.* We then turn to the modelling of a set of brain regions as a *cooperative computation system*--a distributed structure in which each system has its own 'goal structure' for selecting information to act on from its environment, and for transmitting the results to suitable receivers. Finally we sample AI studies of speech understanding. Of particular interest is work by the Carnegie-Mellon Hearsay group on a system using a network of PDP 11's, each functioning as a knowledge source, which interact through a communication center called a *blackboard* rather than being controlled by an executive.

Robot Systems

James S. Albus, and John M. Evans, Jr.
Office of Developmental Automation and Contorol Technology, National Bureau of Standards

Scientific American 234, No. 2, 77-86B, February 1976

Each level in a robot's control hierarchy accepts commands from the next higher level and responds by issuing ordered sequences of commands to the next lower level, making use of sensory feedback to close control loops where they are appropriate. At each level feedback signals are sent to the next higher level and other feedback signals are received from below. These signals indicate the state of the manipulator and the environment. The type of feedback needed depends strongly on the degree of uncertainty encountered in the environment. Large uncer ainties in the robot's world require that the upper levels of control structure incorporate a "world model" which can represent the state of the environment in a meaningful way. When such a robot (e.g. SRI's /Shakey) is faced with an input command to be executed, it tries to set up a hypothetical desired world model and then to devise a set of procedures that can convert the existing world model into the desired one. Various types of industrial robots and experimental approaches to machine vision are briefly discussed.

NEW JOURNAL

CAHIERS DU GROUPE DE TRAVAIL

ANALYSE ET EXPERIMENTATION DANS LES S IENCES DE L HOMME

PAR LES METHODES INFORMATIQUE

No 1, MARCH 1976

E. CHOURAQUI ET J. VIRBEL, ANIMATEURS

CONTENTS

Programme des travaux du groupe de travail

Suites aleatoires selong Kolmogorof et Martin-Lof par Luis Farinas

Informatique et Sciences Humaines: les enseignements d'un colloque par P. Cibois

Nouvelles breves

ADDRESS

Laboratoire d'Informatique pour les Sciences de l'Homme
31, Chemin Joseph-Aiguier
13274 Marseille - Cedex 2
France

The Groupe de Travail is attached to AFCET; Division Theorie et Techniques de l'Informatique.

BIBLIOGRAPHY AND SUBJECT INDEX

CURRENT COMPUTING LITERATURE

Sections: Citations
Author index
Subject index by Computing Reviews categories
Keyword (permuted title) index

Coverage: Reviews and abstracts published in Computing Reviews during 1974 (annual publication)

Address: Association for Computing Machinery
P. O. Box 12105
Church Street Station
New York, New York 10249

Price: $10 for ACM Members
$25 for others

American Journal of Computational Linguistics *Microfiche 50 : 95*

A D M I N I S T R A T I V E D I R E C T O R Y

College and University Computer Science

Author: John W. Hamblen and students

University of Missouri, Rolla

Entries: 300 computer science departments
900 computer centers
societies and government agencies

Contents: Name of chairman or director; address

Degree programs

Computing equipment

Address: ACM Order Department

P. O. Box 12105

Church Street Station

New York, New York 10249

Price: $5 for ACM members and those listed in the Directory

$7.50 for others

100 pages

PRIVACY, SECURITY, AND THE INFORMATION PROCESSING INDUSTRY

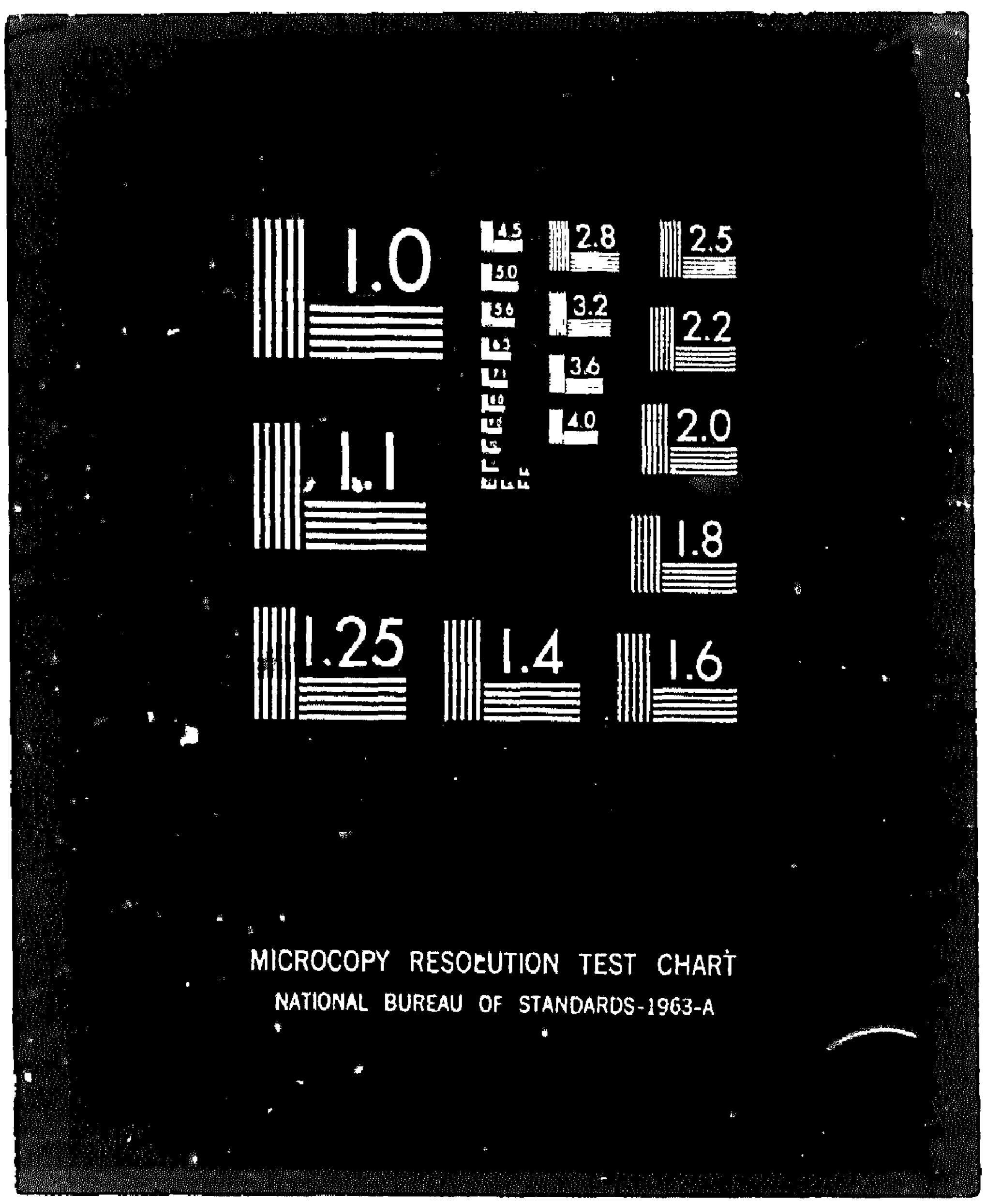
Author: Dahl A. Gerberick, Chairman
Ombudsman Committee on Privacy
Los Angeles Chapter of ACM

Topics: Administrative, technological, and legal considerations
Guidelines for implementation
Comprehensive Right to Privacy Bill
Privacy Act of 1974
Data Center Security Check List
Glossary of Terms
Bibliography

Address: ACM Order Department
P. O. Box 12105
Church Street Station
New York, New York 10249

Price: $9.00 for ACM members
$12.00 for others

END


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