Thrane, T. 1988. Symbolic Representation and Natural Language.

ABSTRACT

The notion of symbolizability is taken as the second requisite of computation (the first being 'algorithmizability'), and it is shown that symbols, *qua* symbols, are not symbolizable. This has farreaching consequences for the computational study of language and for AI-research in language understanding. The representation hypothesis is formulated, and its various assumptions and goals are examined. A research strategy for the computational study of natural language understanding is outlined.

Symbolic Representation and Natural Language
Torben Thrane
Proceedings of NODALIDA 1987, pages 117-154

SYMBOLIC REPRESENTATION AND NATURAL LANGUAGE

by

Torben Thrane
Center for Informatics, University of Copenhagen

1. Introduction: algorithms and symbolic representation

The algorithm is without doubt *the* fundamental concept in computer science. Problems that can be solved by computer are all algorithmic: they can be presented on a form which invites a division of the overall problem into constituent parts, each of which can be solved sequentially and deterministically. When the last constituent problem is solved the overall problem is solved.

To get a computer to solve a problem, however, its constituent parts must be *symbolizable:* they must be put on a form which is accessible to the computer. This precondition is summed up under the rubric 'representation'.

The inescabability of these two preconditions is never in doubt. However, doubts have recently been raised with respect to the value of the comparison that is often made, explicitly or implicitly, between the problem solving capacities of men and machines, and in particular with respect to the role allegedly played by representation as the constitutive feature of those capacities.

Understanding natural language is among the problems whose solution has been expected to be accessible through computer simulation, precisely on the assumption of a common representational basis for the problem solving capacities of men and machines.

2. The representation hypothesis

The topic of the present paper, thus loosely outlined, is *symbolic representation*, in general as well as more specifically with respect to the role it plays in computational linguistics. Initially, the notion of representation can be presented as a simple formula:

(1) **a R b**, which reads: '**a** represents **b**'.

Representation is the name of a relation holding between two entities. The logical properties of the relation are usually taken to be *irreflexivity, intransitivity,* and *asymmetry*. Apart from these logical ones, **R** has properties sometimes summed up by saying that **a** *stands for, complies with, refers to, symbolises, denotes, depicts, designates, corresponds with* **b**.

The logical properties of the entities between which representation holds are more difficult to characterize briefly. Let us assume, initially, that **a** is a physical entity, whereas **b** is typologically unspecified. We return to this issue below.

In relation to (1) we can formulate an overall hypothesis, the *representation hypothesis*, which says:

(2) All intelligent behaviour presupposes the formula (1).

Ultimately, this hypothesis aims to explain how such systems as Miller (1984) called 'informavores', can function as autonomous entities in larger physical environments, which they both affect and are affected by.

2.1. The adequacy of the formula

The formulation (1) invites the view that representation is a contextfree phenomenon, and that it eludes situationally conditioned interpretation. This is incorrect. Already C.S. Peirce - who in this connection can be considered one of the founding fathers of representation theory - insisted on the decisive influence that situation and context has on the interpretation of a sign. And perhaps even more importantly, he insisted that even the recognition of a physical entity as a sign presupposed background, interpretation, and what he called 'semiosis' or - as we shall say - 'the semiotic process': the process by which there is created in an observer a mental correlate - Peirce's 'interpretant' - of a physical phenomenon which, in virtue of this process, now becomes a sign *of* its object, b, to the observer. (*CP* 2.227-9; Hookway, 1985:118-144). From this perspective, nothing is a sign 'in itself'. A sign is *created* - through the semiotic process. So the formula (1) can be amended to:

(3) **a R b** for observer **O** in situation **S** in virtue of the conventions **C**.

This means that an internal representation of **b** has been created in O, 'corresponding to' the physical sign, **a**. This internal representation, according to Peirce, is itself a sign, for which a new interpretant can be created by a recursive application of the semiotic process, and so on, ad infinitum. By way of continuation of the discussion of the logical properties of the entities between which **R** holds, we get a glimpse here of a systematic vacillation in the conception of **a** in the formula: **a** can either be regarded as the *physical sign* which represents **b**; or else **a** can be understood as the *conceptual structure* which has been created in O by virtue of O's taking **a** as a sign for **b**.

If **a** in this way can be thought of as either a physical or a mental phenomenon, **b** must be so conceived as well. It causes no trouble to entertain the idea that physical entities can be represented. Nor does it cause trouble to entertain the idea that mental or abstract phenomena can be represented. There would seem, therefore, to be no trouble in accepting that meaning can be represented, no matter whether meaning is considered to be a mental phenomenon or not; cf. Searle 1983:Ch.8 for a general discussion of this point.

However, there does crop up a problem for computational linguistics. On the view set out above, natural language is itself a representational system of interpreted symbols. At the same time, natural language is - for computational linguistists - a phenomenon that must itself be representable by computationally interpretable symbols. Here the intransitivity of the general representation relation becomes apparent. If it were transitive it would be child's play to construe natural language interfaces, since then, say, a string variable, **a**, would appear to represent whatever the content of **a** represents. But this is not how things work. If **a** in this instance represents anything apart from the sequence of alphanumerical characters that make up the string, then it is a location in the computer's memory, and not, for example, the person that a string 'Tom Jones' is supposed to represent in a given context. From one perspective, then, computational representation of natural language is a special case of hypostasis.

Emerging from this discussion is the following startling fact: symbolic representation of an object, **b**, which *itself* has been interpreted as a symbol, is impossible! It can be a physical entity which - in other circumstances - *could* function as a symbol. Or, it can be a mental phenomenon, an abstract object, a conceptual structure, in addition to whatever meaning is supposed to be.

This conclusion can be summed up in slogan fashion:

(4) *Symbols are not symbolizable*

This slogan may have consequences for the proper study of various linguistic phenomena, anaphora for example. Of more immediate concern, however, is the fact that it can be construed as the basis of the bipartition which characterizes previous attempts to justify the representation hypothesis.

2.2. Stages along the path of the representation hypothesis

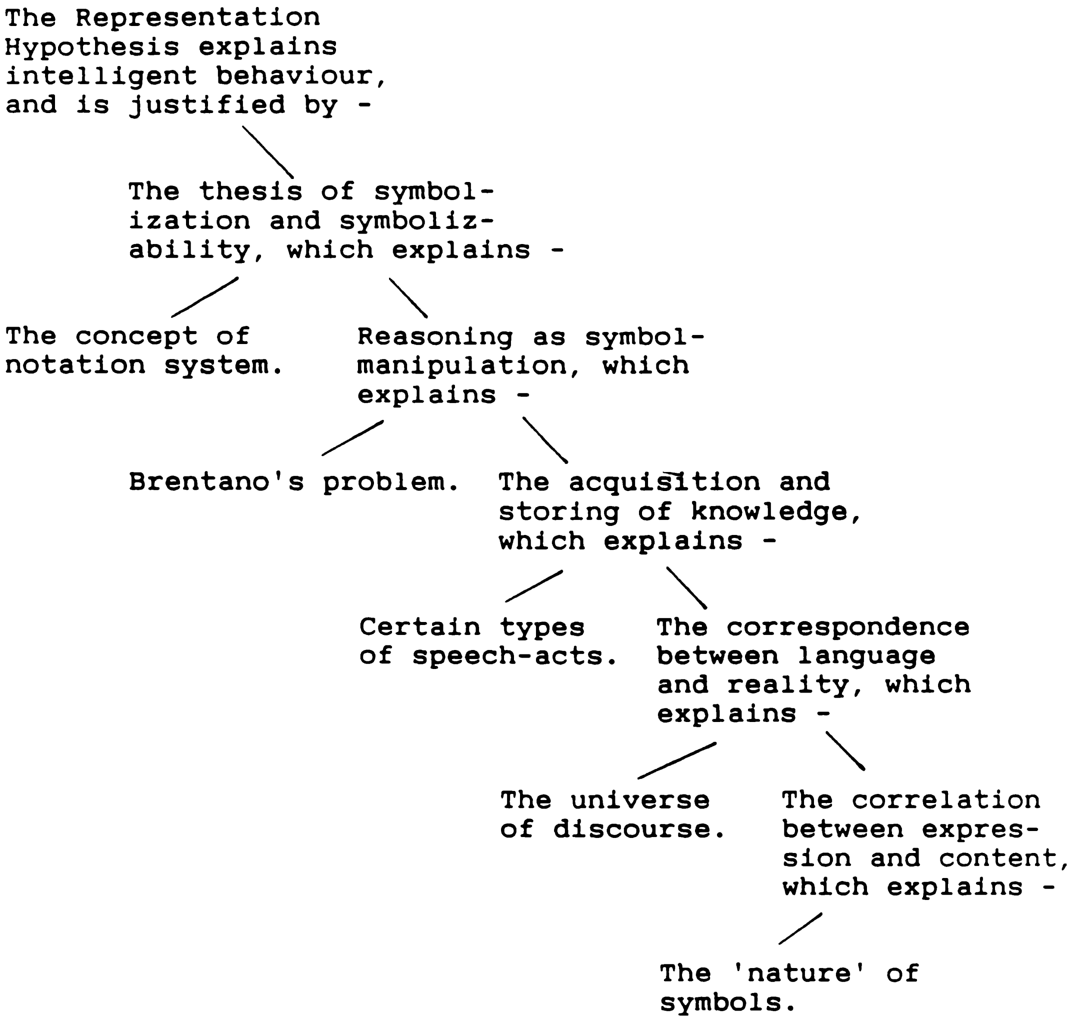
Attempts to justify the representation hypothesis come from many sides and from many more or less related academic disciplines. The same can be said of the attempts to dismiss it as untenable, in particular and most recently by Winograd & Flores (1986).

Let us first dismiss one obvious possibility of discrediting the representation hypothesis. It criticizes any attempt to justify it that takes the form of clarifying (1), correctly claiming that justification of it should take the form of a clarification of (3). Clarification of (1) enters into the ultimate clarification of (3), but they are not the same; nor can clarification of (1) ever on its own count as justification of (2). I shall therefore assume, despite some evidence to the contrary, that everybody who has been seeking justification of (2) in the pursuit of clarification of (1) in fact consciously, if tacitly, have been pursuing clarification of (3).

Clarification of (3) could be regarded as an algorithmic problem for the ultimate solution of which a number of constituent problems have been formulated and described by various academic disciplines and scholarly persuasions.

These constituent problems can with some idealization be presented in a hierarchical structure of mutually explanatory and 'support' systems, as shown in (5):

(5)



The Representation Hypothesis explains intelligent behaviour, and is justified by -

The thesis of symbolization and symbolizability, which explains -

The concept of notation system.

Reasoning as symbol-manipulation, which explains -

Brentano's problem.

The acquisition and storing of knowledge, which explains -

Certain types of speech-acts.

The correspondence between language and reality, which explains -

The universe of discourse.

The correlation between expression and content, which explains -

The 'nature' of symbols.

The right-hand side comprises theories, hypotheses and presuppositions which form essential components of the overall Representation hypothesis, whereas the left-hand side displays (examples of) concrete problems or questions, the answers to which presuppose the validity of the thesis in question. In what follows I will discuss the items on the right-hand side. The items on the left-hand side will be considered in the form of a series of digressions, which together will outline a research strategy for man-machine interaction in natural language.

3. The constituent parts of the representation hypothesis

The list of constituent problems under (5) reads like a catalogue of a significant portion of Western philosophy, cognitive psychology and epistemology in terms of substance. We shall be concerned with them only from the computational point of view.

3.1. Physical symbol systems

The bipartition mentioned above, and the vacillation in the conception of **a** mentioned in 2.1., has led to a bifurcation of computational research which both proceed from Newell's (1980; Newell & Simon 1976) account of a physical symbol system.

A physical symbol system is a system which subscribes to the laws of physics and which at any given time contains a set of structured objects called 'expressions' or 'symbol structures'. Each expression is constituted by a number of *symbols*, ordered in a principled way. Symbols are physical objects whose internal structure may be quite complex, but which have the common property of being combinable with other symbols to form expressions. In addition, a physical symbol system comprises a set of processes that may create, change, destroy, and reproduce expressions. A physical symbol system, then, is a machine which produces a continuous, but continually changing, stream of symbol structures.

The precondition for the proper functioning of a physical symbol system is the notion of *interpretation*, as defined in computer science (Newell 1980:158). Interpretation is there understood as the act of accepting as input an expression which represents a process, and then executing that process.

The first of the two directions of computational research alluded to above studies the internal structure of symbols with a view to establishing a typology of symbols of greater perspicuity and more practical versatility than Peirce's, for example in connection with the generation of fonts and character sets (cf. Knuth 1982; Hofstadter 1982), the creation of MM interfaces, document representation (cf. Levy, Brotsky & Olson 1987; Southall 1987), the development of graphics systems, etc. This direction, under the label 'figural representation', is deeply influenced by the art-aesthetic discussion of the concept of representation, which rejects *similarity* between a and b as the proper grounds for representation in favour of substitution or functional equivalence. Consider in this connection Picasso's famous reply to contemporary criticism that his portrait, *Gertrude*, did not resemble Mrs. Stein: "Don't worry. It will".

The other direction seeks to explain the behaviour of informavores. This direction proceeds from cognitive psychology as much as from semiotics - it is known as 'cognitive science' which subsumes the more practical study of artificial intelligence. This was Newell's own main interest. He formulates the following hypothesis:

(6) *The Physical Symbol System Hypothesis*

> The necessary and sufficient conditions for a physical system to exhibit general intelligent action is that it be a physical symbol system.
>
> Newell, 1980:170

'Necessary' and 'sufficient' in this connection mean, respectively, that a system displaying what we would be prepared to call intelligent behaviour, will always upon closer inspection turn out to be a physical symbol system; and a physical symbol system of sufficient size will always be amenable to organization in such a way that it will display behaviour that we would be prepared to call intelligent.

Clearly, interest in the properties of the symbol differs according to which of the two directions one follows. In the first instance we get an interest in the physical properties of symbols that may alleviate their *extensional* interpretation. In the second, interest centres around the physical properties of symbols that will secure a particular *intensional* interpretation. In these terms the semiotic process can be seen as a complex series of steps whereby a particular physical object undergoes various internal processes (Newell calls them 'symbolic'), whereby they are turned into meaningful structures, *ie* structures that determine the symbol system's subsequent behaviour. It is important to realize that we have no immediate access to these structures, but only recognize them through the behaviour of the system. Consequently, we can only try to describe them on the basis of an abstraction from observed behaviour, in an appropriate formal notation, if the need arises. In this way, both directions take a crucial interest in the physical properties of symbols which converges in the need for interpretation, extensional and intensional. This leads to a digression on notational systems.

3.1.1. Digression: Notational systems

The fundamental property of a symbol is that it is an object manifest to the senses. But - as Newell made clear - a symbol may be of a complex internal structure. This structure will in some cases be amenable to description by means of a notational system, *viz* those cases where the atomic parts of every symbol in the scheme constitute a set that satisfies the five requirements on notational systems formulated by Nelson Goodman (1976):

(7)(a) *Syntactic discreteness:* the decision whether some arbitrary inscription belongs to a particular character and not another is deterministic;

(b) *Syntactic disjointness:* any inscription either belongs to one and only one character, or does not belong to the scheme;

(c) *Unambiguity:* any character, as well as any inscription of any character, must be unambiguous;

(d) *Semantic differentiation:* the decision whether some referent of a particular inscription of any character belongs to one class of objects or another is deterministic;

(e) *Semantic disjointness:* no two characters in a notational system can have any referent in common.

In (8) are displayed samples of signs that are amenable to internal description on the basis of a notational system (a), and of signs the internal structure of which does not reflect a notational system, but which must rather be described on the basis of iconicity (b):

(8)(a)

/* Insert page with figures (8)(a) and (b) and remove this line*/

(b)

The interesting thing about these claims in our connection is that the alphabet satisfies them, whereas larger linguistic entities as a rule do not. The *International Phonetics Association* notation in fact subsumes both types: a subset - used for phonemic transcriptions - forms a notational system on the criteria above, the scheme as a whole - used for phonetic description - does not. Semantic networks, frames, etc., do not constitute notational systems in the required sense, the propositional and predicate calculi do. And finally, all (procedural?) programming languages are notational systems. The parenthesis indicates some hesitation with respect to programming languages like Prolog and Lisp, and many tools specifically developed as system-building aids for knowledge engineering. And the hesitation is due to uncertainty whether to regard such languages from the point of view of the programmer or from the point of view of the machine in making the decision. This ties in with the representation hierarchy (see below, 3.2.1).

3.2. Reasoning as symbolmanipulation

> ..Reason, *when wee reckon it amongst the Faculties of the mind ... is nothing but* Reckoning *(that is Adding and Subtracting) of the Consequences of generall names agreed upon, for the* marking *and* signifying *of our thoughts.*

This passage, from Thomas Hobbes *Leviathan* (1651:I.5), is one of the earliest expressions of what Winograd & Flores somewhat sweepingly style the 'rationalistic tradition' in representation theory. In our own time, Johnson-Laird (1983:2-4) credits Kenneth Craik with the first full-fledged modern version. He describes reasoning as a process that falls into three phases:

(9)(a) A 'translation' of some external process into an internal representation in terms of words, numbers or other symbols;

(b) The derivation of other symbols from them by some sort of inferential process;

(c) A 'retranslation' of these symbols into actions, or at least a recognition of the correspondence between these symbols and external events, as in realizing that a prediction is fulfilled.

Johnson-Laird (1983:2-3)

The symbol has become a mental code with psychological reality, a view which harks back to Peirce's notion of 'interpretant'. However, we should be wary of identifying the two views, mainly because cognitive scientist are more interested than Peirce in explaining the *behaviour* of a system on the basis of representational (semantic) content; cf. Pylyshyn's (1984:39) formulation:

> In cognitive science,..., we want something stronger than derived semantics, *inasmuch as we want to explain the system's behavior by appealing to the content of its representations.*
>
> (my italics)

The justification of this thesis is central to cognitive psychology, for if it can be justified, Brentano's problem disappears.

3.2.1. Digression: Brentano's problem

How can physical stimuli determine the behaviour of a biological system *even though* no direct causal relationship exists between stimulus and behaviour? This is a nutshell formulation of the classical problem which Brentano attempted to solve by introducing a distinction between an 'object' and our mental 'representation' of it, and which, since then, has been one of the constitutive problems of cognitive psychology, on a par with philosophy's mind-body problem.

Rather than solve it, cognitive science believes to have *dissolved* it - with reference to the physical embodiment of symbol systems and the representation hierarchy (Newell 1980:172-75, 1982; Haugeland 1982; Johnson-Laird 1983:399ff; Pylyshyn 1984; but cf. also Winograd & Flores 1986:86-89 and Ch. 8).

A computer can be exhaustively described in various ways: as a physical device, as a collection of logically interpreted circuits, as a deterministic sequence of states defined by a program, etc. The notion of a hierarchy of representational levels stems from the realization that, depending on which type of description is chosen, there is a corresponding, radical change in the proper description of the nature of the information-processing relevant to that type: as electrical current switching on and off, as interpretation of particular electrical patterns as logical values, as interpretation of larger chunks of such patterns as alphabetic characters or numbers, of yet larger chunks as lists of objects or strings of characters, etc.

There is no disagreement on these principles in so far as they are used to describe what *computers* do. But if the account is used as a metaphor - or, indeed, as a literal description - of what *people* do in processing information from physical stimuli of the senses, disagreement is fierce. Searle (1980) draws a useful distinction between 'weak' and 'strong' artificial intelligence research in a discussion of these matters. Weak AI is characterized by regarding computers and programs as tools that enable us to test hypotheses about cognitive processes in a rigorous manner, whereas - in strong AI - the appropriately programmed computer provides the explanation of such processes, with a significant parallel being posited between the jump from hardware to software (in the case of the computer) and from brain to mind (in the case of people). But - to the strong AI researcher (*eg* Pylyshyn) - the point is precisely that it is impossible to define a definite line across which this jump is made, as suggested by the fluid, yet perfectly specifiable, levels of the representation hierarchy.

The representation hypothesis, in conjunction with the strong AI interpretation of the representation hierarchy, provides a sufficiently rich explanation of our interaction with our environment to merit serious consideration. Attempts to discredit it, therefore, must provide a plausible, and equally rich, alternative solution to Brentano's problem to be creditable themselves. This is what Winograd & Flores (1986:Ch. 4) do in their appeal to the cognitive theories of the Chilean biologist Maturana.

3.3. The acquisition and storing of knowledge

The problem of how we *acquire* knowledge is at the epistemological core of Western philosophy. The problem of how we *store* it, once acquired, has only become of prime importance with the advent of computers, for if there is one thing on which all AI researchers agree it is that knowledge is extremely bulky. So, although the question of how best to feed the necessary information into the system is of concern to AI research in general, the burning questions are rather how to cope with its bulk without loss, and how to preserve it in a form suitable for rapid access and retrieval. These problems crucially involve matters of representation, and specific *knowledge representation languages* have been developed to cope with them; cf. Waterman (1986:339-365) for a survey.

All attempts to create knowledge representation languages have assumed the validity of the *knowledge representation hypothsis*, first explicitly formulated by Smith (1982). It goes like this:

(10) Any mechanically embodied intelligent process will be comprised of *structural ingredients* that a) we as *external* observers *naturally* take *to represent a propositional account of the knowledge that the overall process exhibits*, and b) independent of such external semantic attribution, play a *formal but causal* and essential role in engendering *the behaviour that manifests that knowledge.*

Qu.f. Brachman & Levesque (1985:33)
(my italics)

This formulation comprises a series of quite central claims about the nature, organization, and function of the knowledge required for the performance of rational behaviour.

Firstly, knowledge is *representable* - or symbolizable - by *structural* elements. Accessible knowledge is assumed to be organized along previously determined patterns or principles. Only accessible knowledge determines behaviour.

Secondly, the representation of knowledge structured along these lines is assumed to consist of a collection of *propositions*, which we can define here as truth-valued abstractions over states of affairs.

Thirdly, it is supposed to be *natural* for us to see the situation in this light.

Fourthly - and in direct continuation of the digression on Brentano's problem above - it is the knowledge, structured in this way, that is the *cause* of the system's behaviour, and which we call intelligent because it reflects a reasonable 'awareness' of the accessible knowledge. Behaviour is the only external (or interpersonal) evidence of accessible knowledge.

Finally, the triggering of rational behaviour is strictly formal, which means that it is not the propositional *content* as such which is the factor determining behaviour, but rather the structural occurrence of a particular configuration of truth-values in the overall knowledge structure. Pylyshyn's (1984) major aim is to escape this conclusion.

No doubt all of these claims - and their consequences - merit discussion, but I will stick to just one, viz. that acquired knowledge is structured propositionally, and I will do so by way of yet another digression, on speech acts.

3.3.1. Digression: Speech Acts

To make a statement, to ask a question, to issue an order are the three major types of speech act, in the sense that most languages make distinctions in their grammatical systems between the types of sentences typically used to perform them: declarative, interrogative, and imperative, respectively. Among these, declarative sentences have had a particularly prominent position in theoretical discussions of semantics, because they are natural language expressions of streamlined, truth-valued propositions, and because theoretical semantics has often seen it as its major business to account for the conditions that make a particular declarative sentence true.

If, however, the major semantic business is to account for the circumstances in which a particular sentence can be said to have been *understood*, priorities change. Documentation of understanding is *action:* documentation of the understanding of a question is a suitable locutionary act, documentation of an order is execution of a suitable physical act, locutionary or not. These are fairly clear. But what action do we perform in order to document understanding of a declarative sentence?

One possible answer supports the claim above, that knowledge is stored and structured on propositional form. Informally, it says that anyone who hears a declarative sentence will carry through a recursive check as to whether the proposition expressed by it is already part of his 'knowledge base' or not. If it is, and if there is no new evidence for altering one's knowledge on this point, the sentence is dismissed. If it is not, the propositional content of the sentence is checked for *consistency* relative to the knowledge base. If it is consistent, the knowledge base is updated with the new information. If it is not, an assessment is made as to whether a revision of existing knowledge is called for in the light of the new information, or whether the new information is 'wrong', given the validity of the knowledge base. In the former case, the entire knowledge base is revised, including the set of possible inferences that can be drawn from it. In the latter case, the new information is again rejected (or disputed). This, in barest outline, is the mechanism that Johnson-Laird (1983:Ch.15) takes as the foundation of his theory of how we create 'mental models' of our environment.

3.4. The relationship between language and reality

If the representation hypothesis as formulated in (2) is convincing, then its inherent thesis of the relationship between language and world must be too. Winograd & Flores describes the latter thus:

(11) The rationalistic tradition regards language as a system of symbols that are composed into patterns that stand for things in the world. Sentences can represent the world truly or falsely, coherently or incoherently, but their ultimate grounding is in their *correspondence* with the states of affairs they represent. This concept of correspondence can be summarized as:

1. Sentences say things about the world, and can be either true or false;
2. What a sentence says about the world is a function of the words it contains and the structures into which these are combined;
3. The content words of a sentence (such as its nouns, verbs, and adjectives) can be taken as denoting (in the world) objects, properties, relationships, or sets of these.

This is a somewhat simplified account, which I will attempt to show by means of a slightly expanded paraphrase. Sentences are said to *represent* 'states of affairs', because they say something about how the world is organized. If the state of affairs that a sentence represents can be found in the world, then the sentence is true, otherwise false. A state of affairs is a delimited collection of objects which have particular properties and between which particular relations hold. A sentence represents a state of affairs just in case the words or phrases of the sentence, and the structure of which they are a part, refer to those objects that make up the state of affairs, and specify their properties and mutual relations.

No wonder that Winograd & Flores balk at this. The paraphrase implies that there is permanence to a 'true' way in which a sentence represents a state of affairs, that a state of affairs can be identified as the one 'missed' by a particular sentence, that there is an a *priori* global but finite set of objects, properties and relations which we all share a common knowledge of, and which we all have equal (great or small) possibilities of describing 'correctly'. O'Connor (1975) provides further evidence of the insufficiency of the above account of the correspondence theory of truth.

However, a thesis of a *correspondence* relation between language and world need not be tied to such a restrictive formulation. There is nothing in a correspondence thesis that prevents reference to the utterance situation as the constitutive element of linguistic communication. There is nothing to prevent a general account that incorporates speech act theory and correspondence theory. And there is absolutely nothing to prevent us from rejecting the simplistic idea of dependency between states of affairs and linguistic expressions that forms part of the above characterization. This leads to the next digression, on universes of discourse.

3.4.1. Digression: The universe of discourse

Proposition 5.6. in Wittgenstein's *Tractatus* states: *"Die Grenzen meiner Sprache* bedeuten die Grenzen meiner Welt". A possible in-

terpretation of this proposition is that language is what constitutes our world. There are other possible interpretations. I shall provide one more below.
The current interpretation leads to the assumption that the world which language immediately constitutes, is not the actual, physical world, but rather an abstract, a model, or - as we shall call it - a *universe of discourse* or, equivalently, a *discourse universe.*

Discourse universes are private universes, but we can share one or more of them, in part. Long married couples - who are often held to be capable of wordless communication - will, in this jargon, share a large portion of their overall universe of discourse. Universes of discourse may be large and small, and may be more or less densely populated. We can operate on the basis of more than one universe of discourse at the same time, and we can shift from one to another with perfect ease between conversations and even during conversations. And lastly, universes of discourse are intentional, in Searle's (1983:1) sense of 'intentionality':

> Intentionality is that property of many mental states and events by which they are directed at or about or of objects and states of affairs in the world.

The main idea is that language use implies continuous creation, revision, and deletion of universes of discourse for the participants of the conversation. The overall linguistic mechanisms for these purposes are what I have previously styled 'the referential properties of language' (Thrane 1980). Revision and updating of the current discourse universe is in the main associated with identitive, generic and qualitative features (determining definiteness, specificness, certain aspects of conditionality, and referent typology), whereas shifts between universes of discourse are typically associated with presentative and partitive features (determining discourse 'scope', various aspects of definite and indefinite quantification, and other aspects of conditionality). The conditionality of referential expressions is the general mechanism for establishing the 'laws' that hold in a universe of discourse.

On this view, an utterance becomes a *symptom* of the state of the speaker's current universe of discourse, where 'symptom' is intended in the technical sense of Lyons (1977:108) as 'a sign or signal which indicates to the receiver that the sender is in a particular state'. It is incumbent on the receiver, on the basis of his interpretation of the symptom, to gain insight in the state, perhaps to adapt his own universe of discourse accordingly, or to try to persuade the speaker to revise his. Mutual understanding can, under the same view, be regarded as a progressional striving towards the greatest possible congruence between the current discourse universes of speaker and hearer, through cooperation and negotiation, for example about the proper definition or interpretation of a word, or determination of the reference of an expression.

The correspondence relation which is being championed here is a function from a universe of discours to a state of affairs. And accepted truths are those special cases in which the universes of discourse of many (and hopefully, of all) map into the same state of affairs. This does not prevent 'truth' from being the property of just one person - in the sense that the relevant state of his universe of discourse may map into a state of affairs that, over time, will be mapped by the universe of discourse of society at large. This is the only kind of 'absolute truth' that the views taken here will sanction.

3.5. The correlation between expression and content

The last step in this account of the various stages of the representation hypothesis is the question of the connection between expression and content, or meaning. The two currently most favoured bids as to what meaning is, derive from model-theoretic semantics and Situation semantics. Both attempt to characterize meaning in a way that enables them to explain the relation between language and reality, both set off from Frege's distinction between sense (intension) and reference (extension), and both avail themselves of a logical formalism to represent meaning. But from this point on they part company.

Model-theoretic semantics has taken over Leibniz' notion of 'possible worlds', and defines truth relative to that. The intension of a sentence is a function from possible worlds to truth-values, whereas the intension of terms and predicates is a function from possible worlds to, respectively, objects and sets. The extension of a sentence is a truth-value, whereas the extension of terms and predicates is, respectively, objects and sets. So the meaning (intension) of a linguistic expression is those properties of the expression which in any conceivable situation determine what language external objects or sets the expression refers to in that situation, or - if the expression is a sentence - if the expression is true in the situation. More briefly: the intension of a sentence is the set of conditions that has to be satisfied in some possible world for the sentence to be true in that world. Model-theretic semantics is fundamentally intensional, and it represents intensions by means of predicate calculus formulae.

In Situation semantics (Barwise & Perry 1983), meaning is a relation between types of situations. This view stems from the recognition that although there is a principled difference between 'natural' carriers of meaning ('signs'), and 'unnatural' or 'conventional' carriers of meaning ('symbols') - illustrated by the difference between 'smoke means that something is burning' and '"something is burning" means that something is burning' - then both instances involve the transmission of information. Utterance situations, in other words, belong to a type of situation in which the transmission of information is based on symbols and their interpretation. Situation semantics is fundamentally extensional, and its representation of meaning is in fact a representation of situation types, in a formal set theoretic notation.

Even though both model-theoretic and Situation semantics seek to explain the relationship between language and 'the world' through the development of formal notations for the meaning of natural language, both theories suddenly lose sight of language. The model-theoretic solution to the overall problem has the consequence that meaning is divorced from language. If intension is a function from possible worlds (or, in more recent developments, states of affairs indexed for time) to extensions, then the linguistic expression has disappeared and must be reintroduced by a general interpretation function from expressions to intensions. The Situation semantic solution fails to draw what to me appears to be a basic typological distinction between utterance situations (situations *created* by the making of an utterance), and other situations (typically situations *described* by making an utterance). The distinction is based on the notion of intentionality. Utterance situations occur whenever utterances are made; but utterances are the outward manifestation of intentional states of various sorts. Described situations, on the other hand, are not intentional. They are rather 'extentional' in the sense of being the goals at which some intentional states are directed.

Quite apart from such theory-dependent problems, the two semantic theories share a common problem with any other theory that subsumes attempts to represent natural language meaning by means of a formal notation. The creation of any formal representation of the meaning of a natural language sentence amounts to nothing but a claim of synonymy between two material expressions, of course. The problem that must be faced by all semantic theories that rely on formal representations of meaning, is that of interpretability. Even if the formal representation meets all requirements of syntactic well-formedness, internal consistency, etc., it is, in the last resort and in principle, *only* interpretable through the natural language sentence with which it is claimed to be synonymous. The problem of interpretability stems from the non-symbolizability of symbols, *qua* symbols, commented on above (2.1). I take this to be the onus of the second interpretation of Wittgenstein's famous proposition, promised above: my world is only *accessible* through my language. The consequence of this interpretation is that natural language is in fact the *only* efficient meaning representation schema!

3.5.1. Digression: Computational meaning representation

Does this conclusion mean that the representation hypothesis has foundered as a serious explanatary basis for communication, reasoning, and knowledge organization? I don't think so. For even if two well-merited and well-developed semantic theories face problems with respect to their capacity for giving a global characterization of the meaning of natural language sentences, both have developed techniques and insights that can be brought to bear on concrete projects that aim at exploring man-machine communication in natural language. This has never been the primary motivation for model-theoretic semantics nor for Situation semantics.

Such a narrower approach to the problem will have to set off from a set of assumptions specifically geared to, and delimiting, its scope.

First of all, a distinction parallel to Searle's distinction between 'weak' and 'strong' artificial intelligence is called for within the computational study of language. 'Weak', or 'objective' computational linguistics is characterized by

- regarding language as an object of study;
- regarding the computer as a tool;
- regarding a program as a hypothesis of language structure.

'Strong', or 'subjective', computational linguistics, in contrast, is characterized by

- regarding language as a medium of communication;
- regarding the computer as a partner in communication;
- regarding a program as a hypothesis of communicative interaction.

The ultimate goal of the project, on this distinction, falls within 'strong' computational linguistics - and it is still a moot point, to what detailed extent, and to what heuristic level, 'weak' computational methods should enter into it (nature of the parsing required, level of morphological refinement, etc.).

Secondly, it is a limitation that computers, on the 'strong' view above, meaningfully enter into utterance situations only, indeed, utterance situations of an impoverished kind: there can (so far) be no reliance on suprasegmentals, no reliance on gestural or facial information, there is a limited range of language functions, etc. Thus, 'addressing' a computer is, invariably, an attempt to gain access to its universe of discourse, either with a view to changing it or to get documentation of its current state in the form of an appropriate responsive action. If this action is to be based on the computer's 'understanding' of the meaning of a natural language input, meaning in this context is a function from an utterance situation into a discourse universe. This assumption trades on a combination of the functional and relational views of meaning characterizing, respectively, model-theoretic and Situation semantics. It further enhances referential and conative meaning, but deliberately leaves out of account aspects of emotive, phatic, metalingual, and poetic meaning.

Thirdly, the process of 'understanding' is further segmented into a process of 'deciphering' and one of 'interpreting', in the following way: 'deciphering' is a function from an *utterance* to a universe of discourse on the grounds of the 'code', whereas 'interpretation' is a function from one universe of discourse to another. 'Decipherment' will yield a rough approximation to what the hearer takes as the speaker's current universe of discourse, 'interpretation' will work on this to yield a universe of discourse more finegrained, and reflecting the hearer's conception of what the speaker 'has in mind'. Matters of poorly understood words or phrases will be dealt with by 'decipherment', matters of failing reference, inconsistency, etc. by 'interpretation'. 'Interpretation' in this framework is closely akin to the computational view of it (above, 3.1.), in that it involves one or more processes to be executed by the content of a universe of discourse.

Finally, as already mentioned, extension is regarded as a function from discourse universes to *described* situations. Whether the computer in this connection can be said to possess a 'true' image of the world is a question which is in principle no different from the question whether *we* can: it depends on whether our universe of discourse maps into a factual situation. And this in turn depends on the nature, quantity and quality of the *knowledge* we had access to during its establishment.

4. Final remarks

It has not been my primary concern in this paper to try to falsify the representation hypothesis, which I personally find attractive. It has been my concern, on the other hand, to present a series of aspects, interpretations, and consequences of it which in due course may *make* it falsifiable. For if it turns out that the more restrictive semantic program outlined throughout the last three digressions can be carried through, *without* any indication that the same semantic processes characterize interpersonal communication, then there is reason to believe that the representation hypothesis as formulated in (2) is either too strict or wrong. However, one of the results may well be acceptance of the view that the 'nature' of symbols is to be found among the class of topics of which Wittgenstein said:

> *Wovon man nicht sprechen kann, darüber muss man schweigen.*

One of the fascinations about natural language, however, is that it is extremely difficult not to use it, even for discussion of topics that can't be discussed.

REFERENCES

Barwise, J. & Perry, J. 1983. *Situations and Attitudes.* MIT: Cambr. Mass. 2nd Printing 1984.

Brachman, R.J. & Levesque, H.J. (eds.) 1985. *Readings in Knowledge Representation.* Morgan Kaufmann: Los Altos.

CP = *The Collected Papers of Charles Sanders Peirce.* Edited by Charles Hartshorne and Paul Weiss. Cambr. Mass. 1935-66. (References in text are to Volume and Paragraph)

Deuchar, M. 1984. *British Sign Language.* Routledge & Kegan Paul: London.

Goodman, N. 1975. *Languages of Art.* MIT: Cambr. Mass.

Haugeland, J. 1981. Semantic Engines: An Introduction to Mind Design. In Haugeland, J. (ed.) *Mind Design.* MIT: Cambr. Mass. 3rd Printing 1985, pp. 1-34.

Hobbes. T. 1651. *Leviathan.* Pelican Books: Harmondsworth 1963.

Hofstadter, D.R. 1982. Metafont, Metamathematics, and Metaphysics: Comments on Donald Knuth's Article "The Concept of a Meta-Font". In Hofstadter, D.R. *Metamagical Themas: Questing for the Essence of Mind and Pattern.* Bantham Books 1986, pp. 260-287.

Hookway, C. 1985. *Peirce*. Routledge & Kegan Paul: London.

Johnson-Laird, P. 1983. *Mental Models*. Harvard UP: Boston.

Knuth, D. 1982. The Concept of a Meta-Font. *Visible Language* 16, 3-27.

Levy, D.M., Brotsky, D.C. & Olson, K.R. 1987. Formalizing the Figural: Aspects of a Foundation for Document Manipulation. (Draft). Intelligent Systems Laboratory, Xerox Palo Alto Research Centre. 3333 Coyote Hill Road. Palo Alto, Ca. 94304.

Lyons, J. 1977. Semantics. Cambridge UP: Cambridge.

Newell, A. 1980. Physical Symbol Systems. *Cognitive Science* 4, 135-183.

Newell, A. 1982. The Knowledge Level. *Artificial Intelligence* 18, 87-127.

Newell, A. & Simon, H.A. 1976. Computer Science as Empirical Inquiry: Symbols and Search. In Haugeland, J. (ed.) *Mind Design*. MIT: Cambr. Mass., 3rd Printing 1985, pp. 35-66.

O'Connor, D.J. 1975. *The correspondence theory of truth*. Hutchinson: London.

Pylyshyn, Z.W. 1984. *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT: Cambr. Mass., 2nd Edition 1985.

Searle, J. 1980. Minds, Brains, and Programs. In Haugeland, J. (ed.) *Mind Design*. MIT: Cambr. Mass., 3rd Printing 1985, pp. 35-66.

Searle, J. 1983. *Intentionality*. Cambridge UP: Cambridge, 5th Printing 1987.

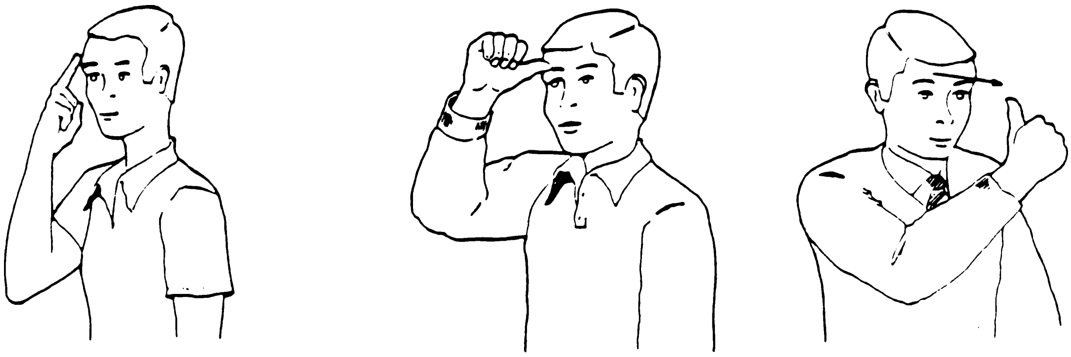
Southall, R. 1987. A basis for the description of machine-written documents. (Draft). Intelligent Systems Laboratory, Xerox Palo Alto Research Centre. 3333 Coyote Hill Road. Palo Alto, Ca. 94304.

Thrane, Torben 1980. *Referential-semantic analysis: Aspects of a theory of linguistic reference*. Cambridge UP: Cambridge.

Waterman, D.T. 1986. *A Guide to Expert Systems*. Addison-Wesley: Reading, Ma.

Winograd, T. & Flores, F. 1986. *Understanding Computers and Cognition. A New Foundation for Design*. Ablex: Norwood, N.J.

Wittgenstein, Ludwig 1921. *Tractatus Logico-Philosophicus*. Routledge & Kegan Paul: London, 1966.



THINK
tab: forehead
dez: index finger extended from closed fist
sig: contact with tab

KNOW
tab: forehead
dez: thumb extended from closed fist
sig: contact with tab

CLEVER
tab: forehead
dez: thumb from closed fist
sig: movement from right to left in contact with tab

Fig. 8a *The structure of BSL signs*

FROM Deuchar (1984)

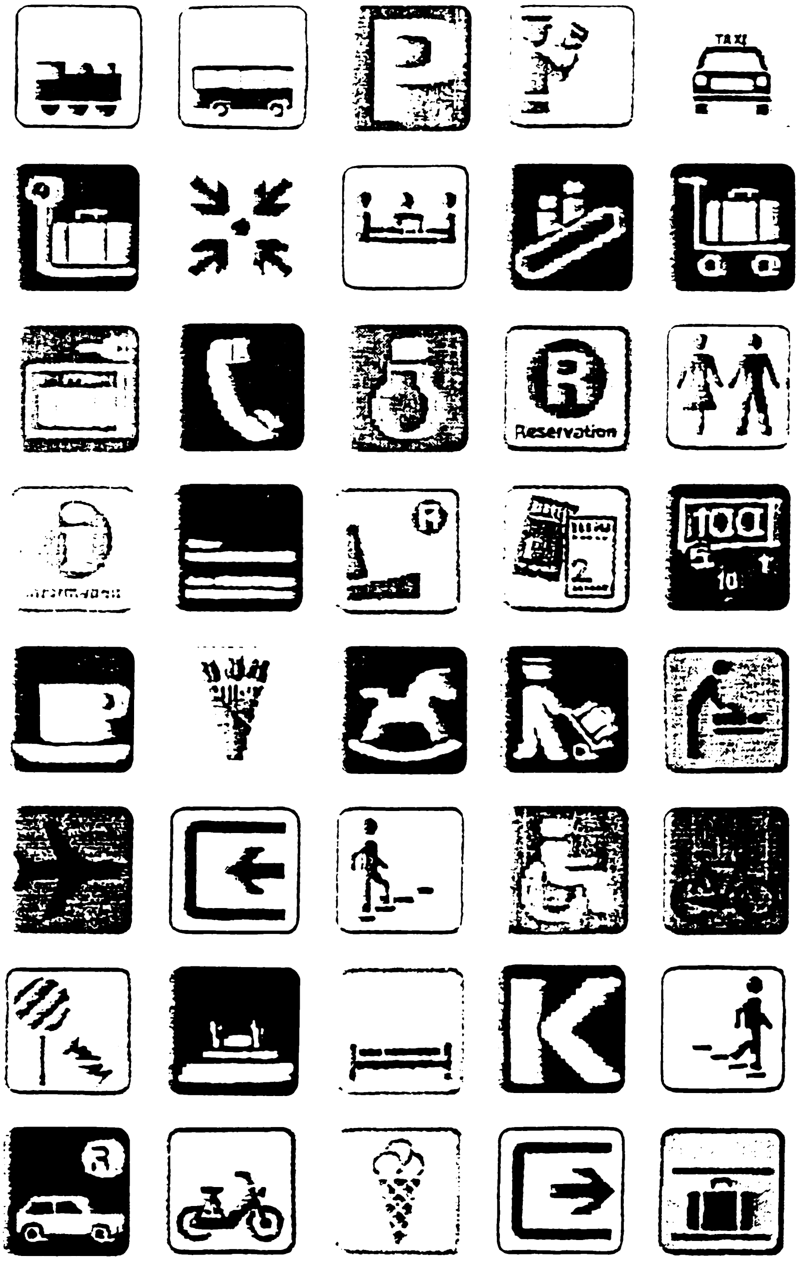


Fig. 8b Icons

FROM DSB Corporate Identity 1990.