Natural Language Properties and Information Systems

E.S.C. van de Ven, P. van Bommel, P.J.M. Frederiks, Th.P. van der Weide

Department of Information Systems, University of Nijmegen
Toernooiveld 1, NL-6525 ED Nijmegen, The Netherlands
{pvb,paulf,tvdw}@cs.kun.nl

Technical Note N9608, October 1996

Abstract

In this paper[1] the concept of natural language is discussed from three perspectives: which *properties* does natural language have from a psychologic point of view, how are natural language sentences *built*, and what is their *meaning*?

The paper serves as a basis for applying natural language in information systems engineering.

Contents

1 Introduction **2**

1.1 Context 2

1.2 Overview 2

2 Properties **3**

2.1 Language is structured 3

2.2 Language is meaningful 3

2.3 Language is referential 4

3 Syntax **5**

3.1 Syntactical categories 6

3.2 Syntactical rules 7

3.2.1 Combining syntactical categories 7

3.2.2 Noun phrases and verb phrases 8

3.3 Transformation rules 8

4 Semantics **10**

4.1 Meaning of words 10

4.1.1 Definitional theory of meaning 10

4.1.2 Prototype theory of meaning 11

4.2 Meaning of sentences 11

5 Conclusions **12**

[1] This paper is based on a chapter of [Ven96].

1 Introduction

1.1 Context

The application of *natural language* has become an important research area in computing science in general, and in information systems engineering in particular. This paper discusses the main aspects of natural language, and explains the context in which natural language is applied in the Department of Information Systems, University of Nijmegen.

A first line to be mentioned focusses on information systems based on *structured information* (e.g. [Hal95]). These systems require information analysis, which is often based on natural language (see also [Win90], [Kri94]). In addition, when validating an information model resulting from the analysis, natural language sentences can be generated (e.g. [DFW96], [FKW96]). This is also called *paraphrasing* of information models.

A second line focusses on information systems based on *unstructured information*, for example in the form of documents. This line is called Information Retrieval ([Rij75]). New techniques in Information Retrieval use document characterizations in which linguistic aspects are incorporated (see e.g. [Sme92]). The growing popularity of the World Wide Web makes this even more urgent. In the information filtering project *Profile* document characterizations are defined in terms of noun phrases ([HSB+96]). Also, the use of semantic relations such as synonyms, antonyms, hypernyms, meronyms, and homonyms is a promising direction here ([BB96]). However, though several attempts to apply linguistics in Information Retrieval have been made, a significant performance improvement still has to be demonstrated.

1.2 Overview

In [BH94] it is argued that a sound definition of the term *language* covering all its necessary and sufficient aspects can not be given. Although linguistics has an empirical nature, it can only be defined as that which is studied by linguistics. This definition of language clearly is unsatisfactory, but alternative definitions are often incorrect or incomplete, take e.g. the following attempt ([Gle91]):

language the sounds produced by the vocal tract which have a meaning and are used to communicate.

Note however that language is not always produced by the vocal tract (writing) and language is not always used to communicate (muttering).

When studying natural language a clear separation between *competence* and *performance* of natural language has to be made. Competence is the knowledge a human being possesses about his mother tongue. This can be seen as a description of language as a system of *signs* and *meanings*. The performance, on the other hand, is about *using* that competence by talking or writing natural language. Although the knowledge about natural language is reflected in the use of it, language is sometimes used in such a way that it conflicts with our knowledge of how it *should* be used. For example, during a conversation we restart, correct or even do not complete sentences.

The systematic part of language has been studied intensively by Noam Chomsky. Chomsky focussed his attention on the systematic part, i.e. competence, of natural language which he used as the basis of his *transformational-generative grammar (TGG)* ([Fah91]). The distinction between competence and performance is also mentioned by the Swedish linguist Ferdinand De Saussure, who stressed that the 'langue' (read competence or the language system) should be separated from the 'parole' (read performance or the way language is used). The 'langue' should be the subject of scientific analysis and should be studied by linguists. The 'parole', however, should be studied by psychologists and sociologists ([Mil91], [Fah91]).

In this paper our attention is focussed on competence. In the discussion of language, English will be used as our main example. But it is important to keep in mind that what is said about English generally goes for the other human languages as well. In the following sections the major properties like structure and meaning of natural language are discussed. In section 2 five major properties of natural language are discussed. The properties *structure* and *meaning*, which are important for information systems engineering, are treated in sections 3 and 4, respectively.

2 Properties

In [Gle91] five major properties of natural language are mentioned, i.e. language is

1. *creative*,
2. *interpersonal*,
3. *structured*,
4. *meaningful*, and
5. *referential*.

The property of creativity states that humans are able to utter and understand sentences they have never heard before. Therefore using language is not a kind of memorization that performs speech acts whenever the appropriate circumstances arise: language is a creative process.

The next property, inter-personality, states that using language is mostly a social activity in which the thoughts of one mind are conveyed to another. The last three properties mentioned, will be discussed in more detail in the following sections.

2.1 Language is structured

Although using language is a creative process, it is also restricted: there are unlimited numbers of strings of English words that we would *not* utter. In information systems terminology this corresponds with the notion of *information structure* (see e.g. [Hal95], [HW93]). Further unwanted sentences may be excluded explicitly using *integrity constraints* ([HW93]) and an additional constraint language, such as Lisa-D ([HPW93]).

As an example of an unwanted sentence, we do not say:

The throws John ball

Our utterances are conform the abstract principles of the language we use. The *structural principles* define how we can combine words into meaningful and comprehensible sentences. In general we are not aware of using these principles. Even so, these principles determine our use of language and allow us to compose and understand boundless new sentences. This structural part of language seems to be very systematic and therefore seems to be a good candidate for formalization. In section 3 more details and examples of the structure of natural language are provided.

2.2 Language is meaningful

Traditionally natural language can be divided into:

- *vocabulary*, consisting of all the words of a natural language.

- *grammar rules*, or *structural principles*, describing the rules that are consulted to construct sentences by combining the words from the vocabulary.

Although this suggests major emphasis on structure, usually semantic aspects are also incorporated here. This is for example the case in *lexicons*, as used in [MRF+90], [BRC93]. In the area of Information Retrieval, lexicons are used for obtaining advanced characterizations of documents (see e.g. [BB96], [Voo94]).

Each word in the vocabulary represents a meaningful *idea* (also called *concept*) about something (e.g. ball), action (e.g. throw), abstraction (e.g. justice), quality (e.g. heavy), etc. In general the relation between word and concept is defined arbitrarily. The purpose of language is to express all the meanings of our utterances (references to concepts) to others, so we have no choice but to learn and memorize these relations.

But people talk in sentences rather than just one word at a time, and the structuring of words, i.e. composing sentences, form a major contribution to the meaning of our utterances. For example the words 'John', 'Paul' and 'killed', can be combined in two different sentences with very different meanings.

John killed Paul.
Paul killed John.

Keep in mind that the difference in meaning between the above two sentences is a result of different ways of structuring the same words. Handling meaning of sentences, will be discussed in more detail in section 4.

2.3 Language is referential

Besides the fact that we know how to put words into meaningful sentences, we also know which words refer to which things, scenes and events in the real world around us, also called *Universe of Discourse* ([Gri82]).

Take as an example the sentence:

That's a rabbit.

The words are put together into a sentence in a correct and meaningful way. However, when this sentence was uttered by a little boy while pointing at a dog, we would not think he has learned English very effectively. This problem is called the problem of *reference*: how to use language to describe the world of real things and events. This problem is quite complex and belongs to the field of psychologists.

In information modelling, the term *reference* is often used in a slightly different way. A distinction is made between concrete (lexical) objects and abstract (non-lexical) objects, also called labels and entities, respectively. A relationship between a label and an entity is called a reference ([Hal95]) or bridge ([Win90]). A relationship between entities was originally called an *idea*, resulting in a Reference and IDea Language, RIDL for short ([Mee82], [DMP84]).

The discussion on meaning and reference of words (and sentences) has also been studied by classical philosophers. But the modern history of the philosophical discussion on meaning has been started by John Locke.

John Locke (1632-1704), an English philosopher, described in his *Essay Concerning Human Understanding*, published in 1690, that words are used to represent the ideas of the speaker: the meaning of a word is the idea that a speaker has in mind when he uses the word and the idea, the

listener creates mentally when he hears the word. These ideas arise from perceptions. So to his opinion meaning and reference are each others equals. This theory can be successfully applied on words that refer to concrete objects in our real world. However, it does not hold for words that refer to *abstract* entities like 'justice' and 'conclusion'.

In the twentieth century the meaning of natural language was greatly influenced by mathematical logic. In the article *Uber Sinn und Bedeutung*, published by Gottlob Frege (1848-1925) in 1892, the distinction between reference and meaning was made clear. Frege, a German mathematician and linguistic philosopher, separated the *reference* of a word (also called *denotation* or *extension*) from the *meaning*[2] of a word (also called *intension*).

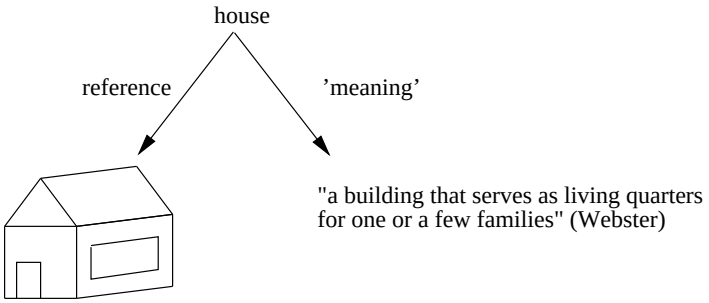

Figure 1: The meaning is separated from the reference

Note that in this theory two expressions with different meanings can refer to the same reference-object. Take e.g. the following two expressions:

George Washington
the first president of the united states

Note that the problem of words referring to abstract entities is unlinked from the meaning but still remains unsolved in the view of some philosophers. The problem of reference is an important issue in the field of psychology and philosophy but is beyond the scope of this paper. Our attention will be focussed on the structure and the meaning (semantics) of natural language.

3 Syntax

In this section the structure of natural language will be investigated and different abstraction levels are distinguished. The structure of languages has been studied intensively in the area of *programming languages* (see e.g. [ASU86]). Focus has been on the following three aspects: grammars, parsing techniques (e.g. [Ned94]), and compilers. Recently, these experiences were applied to natural languages ([KE96]). Of course information systems engineering has also benefitted from the results gained here.

Sentences produced by *natural language users* are built by grouping phrases. These phrases on their turn are composed of words, and these words can also be split in smaller units called *morphemes*. A morpheme is a sequence of *phonemes*. This hierarchy is depicted in figure 2 (adopted from [Gle91]).

Phonemes are described by symbols from the phonetic alphabet. With about 40 phonemes, the 80,000 or so morphemes, and the hundreds of thousands of words can be constructed. The study of phonemes is called *phonology* and the study of morphemes *morphology*. The term *syntax* (i.e.

[2] The meaning is given by means of a definition. In section 4 other approaches to the problem of representation of meaning are presented. The problem of meaning is still a research issue.

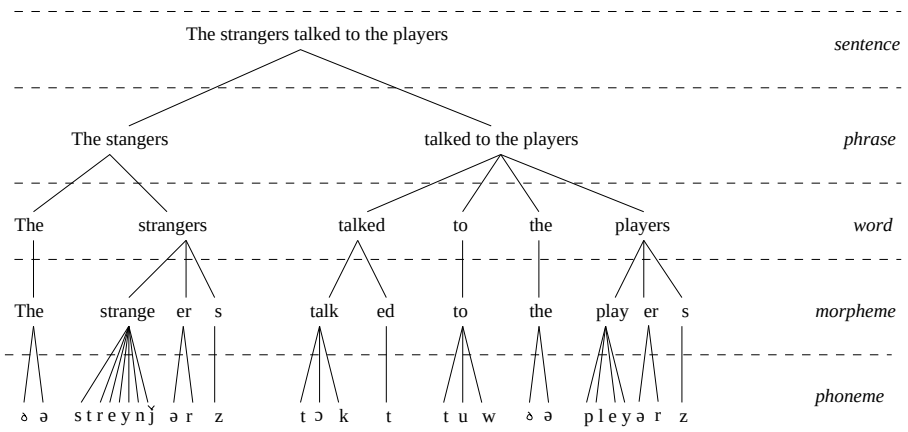

Figure 2: The hierarchy of linguistic structures

'arranging together' (Greek)) is the name of the system that arranges (or groups) words together into meaningful sentences. This topic has been investigated intensively by the American linguist, Noam Chomsky.

As stated before, syntax of natural language consists of two parts:

- *syntactical categories* i.e. the word families. All elements from a certain family can be exchanged with each other without changing the structure of the sentence.
- *syntactical rules* describing the possible arrangements of the syntactical categories.

3.1 Syntactical categories

Traditionally ten syntactical categories can be distinguished ([Mil91]): *noun*, *adjective*, *verb*, *adverb*, *pronoun*, *determiner*, *preposition*, *conjunction*, *numeral*, and *interjection*. In the stepwise procedure (way of working) of the conceptual data modelling technique NIAM ([Hal95]) the syntactical categories *noun* and *verb* play a major role. Nouns usually refer to object types, whereas verbs refer to roles played by object types in fact types. A similar approach to PSM ([HW93]) is presented in [CW93].

A category can be divided into sub-categories; these sub-categories can be divided into sub-subcategories again. Take for example the category of nouns. The noun can be divided into *proper nouns* (e.g. 'John', 'Yoko', 'The Beatles') and *common nouns*. The common nouns consist of *countable nouns* (book, table) and *non-countable nouns* (milk, grain, confidence).

Another example is the category of verbs. The verb-category can be split into three subcategories:

1. *lexical verb* (to hear, to see, to register),
2. *auxiliary verb* (to have, to be),
3. and verbs expressing *modality* (can, must, may, shall, will).

The lexical verbs can be divided again into *intransitive*, *transitive*, *pseudo-transitive*, and *ditransitive* verbs. An intransitive verb (e.g. to sit, to sleep, to smile, and to talk) can not be combined with a direct object. For example we do not say[3]:

[3] A $\star$ indicates the 'invalidity' of the sentence.

⋆(1) *John sleeps the bed.*
⋆(2) *Mary smiles the dog.*

Transitive verb (e.g. to build, to adore, and to devour) however need a direct object. See the examples below:

(3) *John adores his mother-in-law.*
⋆(4) *Mary built.*

Sometimes verbs belong to both categories mentioned above. These verbs *can* have a direct object, but this is not necessary. This category of verbs is called the pseudo-transitive category and contains verbs like 'to eat', 'to drink' and to read'. This situation is shown by the examples given below:

(5) *Peter drinks his milks.*
(6) *Peter drinks.*
(7) *Anne reads a book.*
(8) *Anne reads.*

The last category of verbs is called ditransitive and contains verbs like 'to sell' and 'to give'. The verbs in this category can have both a direct and indirect object.

(9) *John sells the car.*
(10) *John sells the car to Mary.*
(11) *Anne gives the book to Peter.*

3.2 Syntactical rules

3.2.1 Combining syntactical categories

Now, different categories for words have been defined, it needs to be studied how words and categories are related to each other. In conceptual data modelling techniques such as NIAM and KISS ([Kri94]), syntactic categories are related to each other and are represented via graphical models. These models represent so-called *structure sentences*, which are a restricted form of natural language sentences.

Obviously a category contains one or more words. On the other hand, a word can belong to more than one category, i.e. categories are not necessarily disjoint. A word that belongs to more than one category is a hot topic of discussion. In English, it occurs very often that a word belongs to more than one category (e.g. house, work, play, back, paper, surface, etc.). Some believe that such a word should be seen as a single word, while others say that we have to do with different words. The word 'house' as a noun has a phonetic representation that slightly differs from the word 'house' as a verb.

In written language however this difference can not be observed and may lead to confusions. Take for example the sentence below (adopted from [Gle91]). This sentence has two possible interpretations because both words ('bottle' and 'smell') belong to two different categories, namely noun and verb.

(12) *The French bottle smells.*

When 'bottle' comes out as a verb, the sentence is telling us something about what the French put into bottles - namely smells (i.e. perfumes). In another interpretation, the sentence is telling us something about the French bottles.

A sentence that can be interpreted in more than one way, is called *ambiguous*. In the following lines a description of the syntactical rules is provided. With these rules ambiguity can be defined formally.

As mentioned before, syntactical rules describe how syntactical categories can be combined into sentences. Noam Chomsky has researched this intensively, and a lot of theories are based on his results ([BH94]). In [Cho65], Chomsky started the syntactical research. This has resulted in a set of rules that describe how to combine the different syntactical categories. These rules can be expressed by so-called production rules of a context-free grammar ([ASU86]).

3.2.2 Noun phrases and verb phrases

The combination of syntactical categories is important in Information Retrieval in the following context. The characterization of documents can be based on such combinations in order to overcome the simple description via keywords only. It has been proposed to use Noun Phrases as basic building blocks for document characterization, user profiles and queries ([HSB+96]).

As an example of combining syntactical categories, a *sentence* S can be seen as a sequence of a *noun phrase* NP and a *verb phrase* VP.

(SR1) $\langle S \rangle : \langle NP \rangle \langle VP \rangle$

A noun phrase consists of a noun N, that can be preceded by a *determiner* Det and followed by a *prepositional phrase* PP. A noun can be preceded by an unlimited number of *adjectives* Adj. A prepositional phrase is a *preposition* P followed by a noun phrase. In the following grammar, optional arguments are denoted between brackets.

(SR2) $\langle NP \rangle : [Det] \langle N \rangle [\langle PP \rangle]$
(SR3) $\langle N \rangle : [Adj] \langle N \rangle \mid N$
(SR4) $\langle PP \rangle : P \langle NP \rangle$

Because of the different syntactical subcategories of a verb V, a verb phrase can be constructed in different ways. The intransitive verb has no (in)direct object, the ditransitive can have both a direct and indirect object. In the following lines the necessary syntactical rules for the intransitive, transitive and ditransitive verbs are provided.

(SR5) $\langle VP \rangle : V \mid V \langle NP \rangle \mid V \langle NP \rangle \langle PP \rangle$

With these rules sample sentences can be parsed and generated automatically (see e.g. [DKNZ92]). The above mentioned rules are also called rewrite rules.

A sentence can be represented by a parse tree in which each internal node represents the application of one of the rewrite rules. In figure 3.2.2 the parse tree of the sentence

The neighbor of my brother sells the car to Mary

is depicted. This example is adopted from [BH94]. Ambiguity can now be defined as follows:

A sentence is ambiguous if it has more than one parse tree.

Sentence (12) is indeed ambiguous conform this definition.

3.3 Transformation rules

Although these rewrite rules are a powerful mechanism to analyze and describe sentences from our language, they are not powerful enough to describe sentences like (13b) and (14b)[4].

(13a) *Peter read the book.*
(13b) *Did Peter read the book?*
(14a) *The barbarians destroyed Rome.*
(14b) *Rome was destroyed by the barbarians.*

[4] adopted from [BH94]

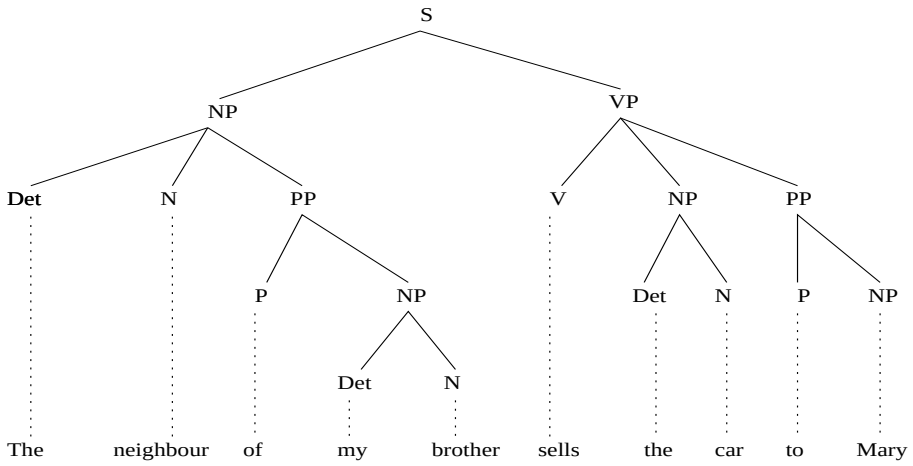

Figure 3: The parse tree of a sample sentence

Chomsky choose not to extend the rewrite rules with rules to produce interrogative (13b) and passive (14b) sentences, but introduced another type of rule: the so-called *transformation rules*. It is clear that in contrast with (13b) and (14b), sentence (13a) and (14a) can be produced by the above mentioned set of rewrite rules. A transformation rule can now be applied to (13a) to produce (13b). This rule transforms a declarative sentence into an interrogative sentence. Another transformation rule is used to transform the active sentence (14a) to the passive sentence (14b). So, to recapitulate, the sentences (13b) and (14b) are produced indirectly, first by the rewrite rules - producing a so-called *deep structure* - secondly by the transformation rules producing the so-called *surface structure* ([BH94]). This situation is depicted in figure 4, which is adopted from [BH94].

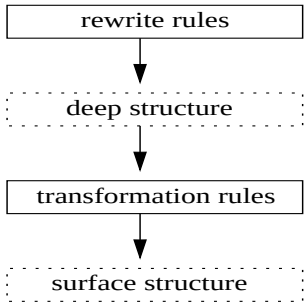

Figure 4: The model of the transformational - generative grammar.

Transformation rules can be defined arbitrarily. A transformation rule can transform everything into everything. Restrictions are needed to distinguish meaningful transformation from senseless transformation rules. The restrictions on these rules can only be obtained by induction on an empirical basis. The last 25 years of research on syntactical aspects of language has been focussed on finding empirical restrictions and on formulating these restrictions adequately. These issues still play an important role in current research.

Rewrite and transformation rules facilitate the description of language and to check sentences on well-formedness on a syntactical level. In the following section it is described how semantical anomalies, i.e. sentences without an explicable meaning (e.g. *The table writes a letter*), can be excluded.

In the field of computer science a lot of experience has been gained with program transformations ([Par90]). Transformations are also important for compiler construction, where this is called *trans-*

duction. The idea behind transduction is to transform 'sentences' in one language to 'sentences' in another language. In this context language should be considered in a wider sense. For example, in information systems engineering, structured sentences are transduced to information models (NIAM and KISS).

4 Semantics

In this section it is described how the meaning of our utterances can be defined and used. A discussion on the meaning of single words is provided, followed by a discussion on meaning of complete sentences.

4.1 Meaning of words

It is generally accepted by linguists that the meaning of a word is a *concept*. So a word is just a representative of a meaningful idea (or concept) in the mind of the person using that word. During communication we try to express these concepts to others. The question,

How to express these concepts?

is one of the most difficult questions in the study of language. The subfield that attempts to answer this question is called the study of *semantics*.

Currently words and their meaning are extensively studied, resulting e.g. in lexica such as WordNet ([MRF+90]), which can be used for several directions in informations system engineering research (see e.g. [BR96]). Lexica provide the user with the possibility to search for e.g. synonyms, antonyms, hypernyms, and homonyms.

As stated before, one of the oldest approaches to the topic of meaning equates concepts of words and phrases with reference. We have already seen that this theory of meaning results in several difficulties. Therefore two other approaches are defined: the *definitional theory of meaning* and the *prototype theory of meaning*.

4.1.1 Definitional theory of meaning

In figure 1 the word *meaning* is placed between quotation marks, because - as just stated - the meaning of a word is a concept instead of a definition. A concept is an idea in someone's mind, a definition however is written by the author of a lexicon. It is the definitional theory of meaning that assumes that these two fulfill the same role.

This approach states that meanings can be analyzed into a set of subcomponents, organized in our minds much as they are in standard dictionaries. Various meaning (or semantical) relationships exist between words and phrases. Some words are similar in meaning (*synonym* relationship) and other are opposites (*antonym* relationship). According to this approach these relations can be explained by assuming that words are sets of *semantic features*. The concept of 'bird' for example contains the semantic features 'feathers', 'flies', 'animal' and 'wings'.

Taken together, the semantic features constitute a definition of a word. According, to this theory, we carry such definitions in our heads as the meaning of words in a so-called *mental lexicon* ([Mil91]).

Although this approach is intuitively satisfying, some problems are brought into being: language itself is used to express the meaning of language-elements. To break this vicious circle, philosophers - from Plato to Leibniz - have been trying to make a list of so called axiomatic 'base-words', that can be used to describe the meaning of all other words ([Mil91]). Up to now all these attempts have failed.

4.1.2 Prototype theory of meaning

Another approach that attempts to define concepts is the *prototype theory* of meaning. This theory explains the fact that some members of a meaning category appear to exemplify that category better than others do. The definitional theory, lists the necessary and sufficient semantic features that *define* a concept. However an 'armchair' seems to be a better example of the concept of 'furniture' than a 'reading lamp'. An armchair is a typical piece of furniture, a reading lamp is not. This difference cannot be explained by the definitional theory of meaning. It claims to have said it all by the following definition of 'furniture' (conform Webster's dictionary).

furniture movable articles used in making a room ready for occupancy or use

The prototypical theory states that a concept is defined by a whole set of features, no one of which is individually either necessary or sufficient.

Consider e.g. the concept of 'bird'. Prototype theorists claim that 'birdiness' is determined by the total number of features a given creature exhibits. Animals that have few (penguins and ostriches) will be judged to be pore members of the bird family, while those that have many (such as robins) will seem to be a strong member (or prototype).

It appears that both the definitional and the prototypical approaches to word meaning have something to offer. To prototype theory explains why a robin is a 'better' bird than an 'ostrich', and the definitional theory says that an ostrich should nevertheless be classified as a bird (conform Webster's dictionary).

bird any of a class of warm-blooded vertebrates distinguished by having the body more or less completely covered with feathers and the forelimbs modified as wings

4.2 Meaning of sentences

So far, it has been discussed how to handle the meaning of words. When studying the meaning of sentences we come to a problem at a higher level of complexity. It has been shown that the meaning of a single word is a concept which can be 'described' by means of a definition or a prototype. The meaning of a complete sentence on the other hand, has to do with relations *between* these concepts. Evidently, the context sensitive nature of natural language - ignored by the rewrite-rules of the transformational-generative grammar - supply a major contribution to the meaning (i.e. semantics) of natural language.

Basic sentences introduce some concept that they are about (called the *subject* of the sentence). Simon C. Dik mentions ([Dik89]) that the term 'subject' only makes sense when a certain context (i.e. the *predicate* of the sentence) is provided. In a sentence like (15) 'The boy' is called the subject, and what is proposed or predicated of this concept is that he 'hit the ball'.

(15) *The boy hit the ball.*

Generalizing from this example, the meaning of a sentence (sometimes called *proposition*) can be regarded as a sort of miniature drama in which the verb is the action and the nouns are the performers, each playing a different role. In the example above of the *boy-hitting-ball* miniature drama, the 'boy' is the *doer*, 'the ball' the *done-to* and 'hit' is the *action* itself ([Gle91]). This *action-oriented* viewpoint treats verbs as basic frames to be filled with concepts. This approach enables us to describe the meaning of a sentence precisely.

However, the linguistic theory of the transformational-generative grammar, is a formal attempt at structuring syntactical categories in terms of rules of formal syntax to be applied *independently* of the meanings. In this theory syntax is thus given priority over semantics. Of course, semantics

lexical item	smile	eat	sell
category	V	V	V
sub-categorization	<>	$< (NP) >$	$< NP(PP) >$

Table 1: The lexicon contains the context sensitive information

is recognized by this theory, but is not its basic assumption. How this linguistic theory handles semantics will be described shortly.

The transformational-generative theory states that context sensitive information should be stored, separately from the rewrite-rules, in a lexicon. This implies that the context sensitive information describing to which sub-category a verb belongs, should also be stored in the lexicon (see table 1).

Note that this context sensitive information is needed for the benefit of a correct deep-structure generation. Therefor the lexicon is consulted *during* deep-structure generation by choosing the appropriate lexical items to fill the structure. With this information semantic anomalous sentences like,

John walks the house,

can be excluded.

Another form of context sensitive information are the so-called *selection restrictions*. The selection restriction expressed in table 2, states that the PP of 'to talk' should be animate. With this restriction rule, semantic anomalous sentences like,

John talks to the table,

can be excluded.

The linguistic theory, called *functional grammar* ([Dik89]), which deals with semantics, forms the basis for the formal modelling technique Conceptual Prototyping Language CPL ([Dig89]).

lexical item	talk
category	V
sub-categorization	$< (PP_{[animate]}) >$

Table 2: An example of a selection restriction

5 Conclusions

In this paper the concept of natural language has been discussed from the perspectives *syntax* and *semantics*. The paper serves as a basis for applying natural language in information systems engineering, including information systems based on *structured information*, and information systems based on *unstructured information*, for example in the form of documents.

References

[ASU86] A.V. Aho, R. Sethi, and J.D. Ullman. *Compilers: Principles, Techniques, and Tools.* Addison-Wesley, Reading, Massachusetts, 1986.

[BB96] F.C. Berger and P. van Bommel. Augmenting a characterization network with semantical information. *Information Processing & Management*, 1996. (To appear).

[BH94] H. Bennis and T. Hoekstra. *Generative Grammatica.* Floris Publications, Dordrecht, The Netherlands, 1994. (In Dutch).

[BR96] J.F.M. Burg and R.P. van de Riet. COLOR-X: Using Knowledge from WordNet for Conceptual Modeling. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database and Some of its Applications.* MIT Press, 1996.

[BRC93] J.F.M. Burg, R.P. van de Riet, and S.C. Chang. A data-dictionary as a lexicon: An application of linguistics in information systems. In Bhargava B., Finin T., and Yesha Y., editors, *Proceedings of the 2nd International Conference on Information and Knowledge Management*, 1993.

[Cho65] N. Chomsky. *Aspects of the theory of syntax.* MIT Press, Cambridge, Massachusetts, 1965.

[CW93] M.A. Collignon and Th.P. van der Weide. An Information Analysis Method Based on PSM. In G.M. Nijssen, editor, *Proceedings of NIAM-ISDM.* NIAM-GUIDE, September 1993.

[DFW96] C.F. Derksen, P.J.M. Frederiks, and Th.P. van der Weide. Paraphrasing as a Technique to Support Object-Oriented Analysis. Technical Report CSI-R9603, Computing Science Institute, University of Nijmegen, Nijmegen, The Netherlands, January 1996.

[Dig89] F.P.M. Dignum. *A Language for Modelling Knowledge Bases.* PhD thesis, Free University, Amsterdam, The Netherlands, 1989.

[Dik89] S.C. Dik. *The Theory of Functional Grammar. Part I: The Structure of the Clause.* Floris Publications, Dordrecht, The Netherlands, 1989.

[DKNZ92] C. Dekkers, C.H.A. Koster, M.-J. Nederhof, and A. van Zwol. Manual for Grammar WorkBench Version 1.5. Technical Report 92-14, Department of Computer Science, University of Nijmegen, Nijmegen, The Netherlands, July 1992.

[DMP84] O.M.F. De Troyer, R. Meersman, and F. Ponsaert. RIDL User Guide. Research report, International Centre for Information Analysis Services, Control Data Belgium, Inc., Brussels, Belgium, 1984.

[Fah91] C. Fahner. *Taal voor Welzijn.* De Vijverberg, Kampen, The Netherlands, 1991. (In Dutch).

[FKW96] P.J.M. Frederiks, C.H.A. Koster, and Th.P. van der Weide. Validation of Object-Oriented Analysis Models using Informal Language. Technical Report CSI-R9609, Computing Science Institute, University of Nijmegen, Nijmegen, The Netherlands, May 1996.

[Gle91] H. Gleitman. *Psychology.* W.W. Norton and Compagny, New York, New York, 1991.

[Gri82] J.J. van Griethuysen, editor. *Concepts and Terminology for the Conceptual Schema and the Information Base.* Publ. nr. ISO/TC97/SC5-N695, 1982.

[Hal95] T.A. Halpin. *Conceptual Schema and Relational Database Design.* Prentice-Hall, Sydney, Australia, 2nd edition, 1995.

[HPW93] A.H.M. ter Hofstede, H.A. Proper, and Th.P. van der Weide. Formal definition of a conceptual language for the description and manipulation of information models. *Information Systems*, 18(7):489–523, October 1993.

[HSB+96] E. Hoenkamp, L. Schomaker, P. van Bommel, C.H.A. Koster, and Th.P. van der Weide. Profile - A Proactive Information Filter. Technical Note CSI-N9602, Computing Science Institute, University of Nijmegen, Nijmegen, The Netherlands, 1996.

[HW93] A.H.M. ter Hofstede and Th.P. van der Weide. Expressiveness in conceptual data modelling. *Data & Knowledge Engineering*, 10(1):65–100, February 1993.

[KE96] C.H.A. Koster and E. Oltmans (Eds.). Proceedings of the first AGFL Workshop. Technical Report CSI-R9604, Computing Science Institute, University of Nijmegen, Nijmegen, The Netherlands, January 1996.

[Kri94] G. Kristen. *Object Orientation, the KISS Method: From Information Architecture to Information System.* Addison-Wesley, Reading, Massachusetts, 1994.

[Mee82] R. Meersman. The RIDL Conceptual Language. Research report, International Centre for Information Analysis Services, Control Data Belgium, Inc., Brussels, Belgium, 1982.

[Mil91] G.A. Miller. *The science of words.* Scientific American Library, New York, New York, 1991.

[MRF+90] G.A. Miller, Beckwith R., C. Fellbaum, D. Gross, and K.J. Miller. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, 3(4):234–244, 1990.

[Ned94] M.-J. Nederhof. *Linguistic Parsing and Program Transformations.* PhD thesis, University of Nijmegen, Nijmegen, The Netherlands, 1994.

[Par90] H. Partsch. *Specification and Transformation of Programs - a Formal Approach to Software Development.* Springer-Verlag, Berlin, Germany, 1990.

[Rij75] C.J. van Rijsbergen. *Information Retrieval.* Butterworths, London, United Kingdom, 1975.

[Sme92] A.F. Smeaton. Progress in the Application of Natural Language Processing to Information Retrieval Tasks. *The Computer Journal*, 35, June 1992.

[Ven96] E.S.C. van de Ven. A Study on the Profits of Semantics in Object-oriented Information Modelling based on Natural Lanuage – An Artificial Intelligence approach. Master's thesis, University of Nijmegen, June 1996. No. 371.

[Voo94] E.M. Voorhees. Query Expansion Using Lexical-Semantic Relations. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, July 1994.

[Win90] J.J.V.R. Wintraecken. *The NIAM Information Analysis Method: Theory and Practice.* Kluwer, Deventer, The Netherlands, 1990.