Information Extraction:
State-of-the-Art Report

Robert Engels & Bernt Bremdal

CognIT a.s, Asker, Norway

Identifier	nr. 5
Class	**Deliverable**
Version	**2**
Version date	**July 28, 2000**
Status	**Final**
Distribution	**Public**
Responsible Partner	**CognIT a.s**

On-To-Knowledge Consortium

This document is part of a research project funded by the IST Programme of the Commission of the European Communities as project number IST-1999-10132. The partners in this project are: Vrije Universiteit Amsterdam (VU) (co-ordinator), NL; the University of Karlsruhe, Germany; Schweizerische Lebensversicherungs- und Rentenanstalt / Swiss Life, Switzerland; British Telecommunications plc, UK; CognIT a.s, Norway; EnerSearch AB, Sweden; AIdministrator Nederland BV, NL.

Vrije Universiteit Amsterdam (VU)
Faculty of Sciences, Division of Mathematics and
Computer Science
De Boelelaan 1081a
1081 HV Amsterdam, the Netherlands
Fax and Answering machine: +31-(0)20-872 27 22
Mobil phone: +31-(0)6-51850619
Contactperson: Dieter Fensel
E-mail: dieter@cs.vu.nl

University of Karlsruhe
Institute AIFB
Kaiserstr. 12
D-76128 Karlsruhe, Germany
Tel: +49-721-608392
Fax: +49-721-693717
Contactperson: R. Studer
E-mail: studer@aifb.uni-karlsruhe.de

**Schweizerische Lebensversicherungs- und
Rentenanstalt / Swiss Life**
Swiss Life Information Systems Research Group
General Guisan-Quai 40
8022 Zürich, Switzerland
Tel: (41 1) 284 4061, Fax: (41 1)284 6913
Contactperson: Ulrich Reimer
E-mail: Ulrich.Reimer@swisslife.ch

British Telecommunications plc
BT Adastral Park
Martlesham Heath
IP5 3RE Ipswich, UK
Tel: (44 1473)605536, Fax: (44 1473)642459
Contactperson: John Davies
E-mail: John.nj.Davies@bt.com

CognIT a.s
Busterudgt 1.
N-1754 Halden, Norway
Tel: +47 69 1770 44, Fax: +47 669 006 12
Contactperson: Bernt. A.Bremdal
E-mail: bernt@cognit.no

EnerSearch AB
SE 205 09 Malmö, Sweden
Tel: +46 40 25 58 25; Fax: +46 40 611 51 84
Contactperson: Hans Ottosson
E-mail: hans.ottosson@enersearch.se

AIdministrator Nederland BV
Julianaplein 14B
3817CS Amersfoort, NL
Tel: (31-33)4659987, Fax: (31-33)4659987
Contactperson: Jos van der Meer
E-mail: Jos.van.der.Meer@aidministrator.nl

Revision Information

Revision date	Version	Changes
2000-05-25	1	First Draft by CognIT a.s
2000-07-28	2	Final Version

Calvin: "I like to verb words."
Hobbes: "What?"
Calvin: "I take nouns and adjectives and use them as verbs. Remember when `access' was a thing? Now it's something you do . It got verbed."
Calvin: "Verbing weirds language."
Hobbes: "Maybe we can eventually make language a complete impediment to understanding."

- *Calvin & Hobbes, by Bill Watterson*

Table of Contents

1 Theory and Principles; Information, Knowledge and Meaning

1.1 *introduction*

Language, knowledge and our perception of the world are closely related. In order to create a foundation for our theories and to judge historic and present initiatives with respect to information extraction we will briefly discuss some philosophical aspects that address the human ability to perceive the world, experience it and to express a view of the same world. Essentially we will attempt to justify our position for the subsequent evaluation of different historic initiatives to natural language processing (NLP), information retrieval and text understanding. Basically what we claim is that a major set of initiatives in development of knowledge-based systems and NLP have struggled and sometimes failed simply because they have adopted a very classic scientific approach. The success of modern mathematics in natural sciences like physics have reinforced the notion that more mundane and qualitative aspects of the world can be modelled by formalisms that are neutral to both individual and context alike. The use of formal logic and the application of normalised language formalism have created the basis for multiple systems development. Although attractive in terms of organising knowledge and information and to rationalise the processing of such it creates a divide between the real world problems and many of the systems developed. One well known issue debated for years within the AI community is the frame problem which defines a limited and protected world within the systems that can be operated. Once the frame issue is violated the performance of the system falters. The issue then becomes to map the real world onto the artificial through a process that is called world representation. Consequently the advantage gained with strict formalisms is lost through the difficulty of world representation. Several systems have faltered during the scale-up process rendering no practical benefits due to this problem. Another aspect that classic science imposes on system development is the need to make things objective. The issue of replication of use, both with respect to full systems and stand-alone knowledge bases, is intimately related to the desire for objectiveness. Moreover, several systems that we have studied apply a very mechanistic way of analysis mostly following a linear pattern of processing. Again, such a process is true to scientific tradition staked out by Descartes and others. We advocate a more holistic approach to analysis using non-linear approaches that is inspired by nature itself. All this is not to say that science and philosophy do not offer other approaches to the issues at hand. On the contrary, modern physics and art do introduce measures of relativism that we feel are important to acknowledge when working with NLP. We also claim that subjectivity poses no threat to NLP and the possibility of replication. On the contrary it has profound effects on how to appreciate language as a means of addressing knowledge and for communication among people. We also take the stance of key authors in the field of knowledge management (Nonaka and Takeuchi, 95) that claim that articulation of experience is language dependent. Tacitness is a consequence of internalisation through processing of sensory and mind impressions as well as a lack of vocabulary that sufficiently express the concepts internalised.

All of this raises the issue of how to approach computer based natural language processing and information extraction. It influences our way of thinking about ontologies and general epistemology. It helps us to explain the success of recent advances in statistical natural language processing compared to that of traditional NLP. It also helps us to define a theory that can support the pursuit of intelligent text understanding by means of computers.

1.2 *Philosophical basis*

John Locke was the founder of British empiricism. He claimed that things existing in the real world are objective in nature. Even if the sensory perception of things is illusory, it is undoubtedly evident that something can be perceived. People’s sensory perception capabilities are their

primary source of many ideas. There is another fountain, the reflection capabilities of our mind, from which experience furnish the understanding with new ideas.
Gottlob Frege is said to be the father of "analytic philosophy". He built a pure language of mathematics by introducing strict terms and definitions of the numbers. Basically he created a language of modern logic that was without ambiguities. His work departed from established work of Descartes and his like. But he was again a pure rationalist that discarded all aspects of intuition and meta-physics to the extent that this obscured the essence to be communicated. His efforts spun off results that had profound effects on other philosophers' concept of language, basically because he was the first to distinguish between meaning and reference. He claimed that the semantics of a sentence is purely a claim. Its reference in the real world provides its true value. In this manner he was able to explain how people can have a different set of opinions of entities of the real world, without necessarily disputing the real world itself.

Husserl was the father of phenomenology and focused on the relationship between thinking itself and the world. Husserl addresses the human perception of the world and the notion of experience. This gave rise to consciousness of experience and its effect on our ability to perceive the world. It was emphasised by Husserl that intention is the token of all consciousness. The fundamental aspect of intention is the will to act or the action itself. Husserl gave rise to the ideas of another philosopher, Heidegger. Heidegger used Husserl as basis for the notion of "being in the world" (German: "Dasein"). "Dasein" defines relationships between different things in the world. Heidegger proposes then that observation and being is basically inseparable and thus places the observer in the midst of the action. There is no such thing as a detached spectator of the world. Hence Heidegger points to important aspects of subjectiveness related to our perception of the world. Both Husserl and Heidegger are important in order to understand the purpose of both ontologies and data. Given that we accept the view of Heidegger, it is imperative that two people can never gain exactly the same view of the world. In order to share ideas of the same world they need to find ways of sharing their experience of it. Naturally, one way of doing this is through an ordinary dialogue. The challenge is then to map what they have experienced onto the knowledge of the other observer. This challenge is amplified not only by the different impressions of the world, but also the environment that developed the language to articulate at least part of what they experienced. We see that Freges reference model constitutes a basis for this, it also tells us that our experiences form the concepts in our mind (see figure 1). This again forges our ways of communication. Language must be used to bridge experience. But daily language itself contains the type of ambiguities that Frege worked hard to remove from the language of logic. Hence we are faced with three types of mismatches: the *subjectiveness of experience* that yields a type of *subjectiveness in expression* and an expression that uses a *language that is prone with ambiguities and incomplete references*.

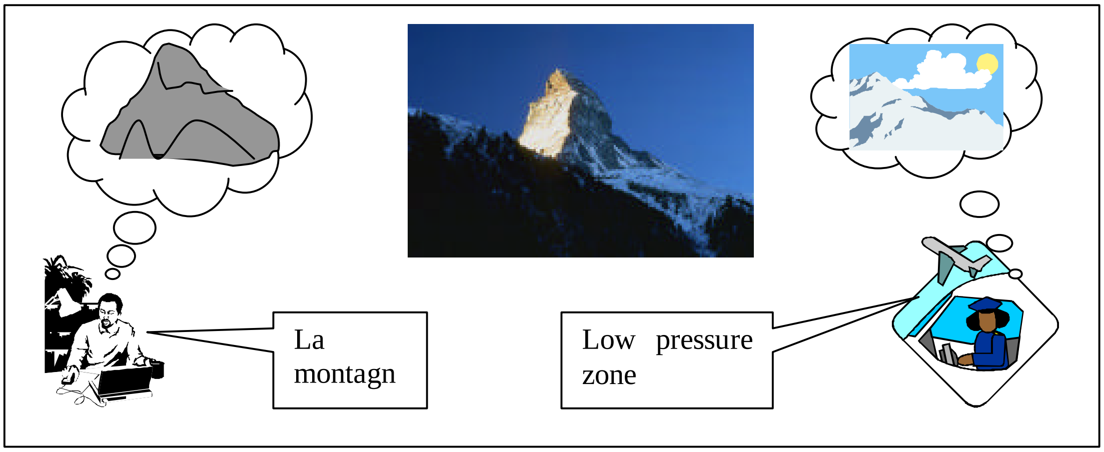


Figure 1: Difference in perception of the world, the author and the pilot

Another modern philosopher was Wittgenstein. His early concerns were tied to the task of describing phenomena. At that time Wittgenstein used logic as formalism for natural language, whereas Frege had already demonstrated the power of a pure formalism for logic. Wittgenstein tried to show that language was as an image of reality that corresponds exactly to logic. During his career Wittgenstein developed his ideas about language and he became firm in his belief that we had to attempt to understand the concepts that language addressed and not the language itself. Clearly there is a distinction between the word and the concept (Wittgenstein 68):

"You say: the point isn't the word, but its meaning, and you think the meaning as a thing of the same kind as the word, though also different from the word. Here the word, there the meaning. The money, and the cow that you can buy with it."

-- *Wittgenstein* --

Wittgenstein clearly expresses concern about words as the holder of a meaning that is faithful to the world concept that it addresses. Clearly this is not so. We know very well that a word can have several different meanings. Wittgenstein was content that the true conception of the world could not necessarily be conveyed by the word alone, but by its use. "For a large class of cases – though not for all – in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language." To us this is key to some of the success that recent advances in statistical natural language. Statistics is a fine way of determining how a word is used.

According to Miller (Miller, 2000) the effort of representing nature has always been a central problem for scientists and artists in the Western world. Again there is evidence that there has been a mutual influence such that the effort and opinions of scientists have been absorbed in arts and vice versa. Miller states that Albrecht Dürer and Leonardo da Vinci, among other Renaissance artists struggled with mathematical and physiological problems concerning linear perspective. In the late nineteenth and early twentieth centuries, artists once again became interested in problems of space and time. Miller tells us that Georges Braque, Paul Cézanne, and Picasso were preoccupied with this during their career. It is an interesting historic issue that Braque and Picasso discovered Cubism in Paris at the same time that Einstein discovered relativity in Bern. This stirred heated discussions in terms of objectiveness and subjectiveness. Western art and science, like philosophy have at times been in dire straits regarding this. In particle physics the issue surfaces in full force with the departure from the Newtonian world image to Einstein's' "new world" order. The same discussion in terms of relativity and observers vantage point can be seen in discussion on art. The concepts of beauty and aesthetics have always been an issue. Millers' objective assessments of aesthetics and beauty attempts to remove subjective elements. The aesthetician Clive Bell (1881-1964) was a British art critic and philosopher of art who defended abstract art. Bell's aesthetic theory was focused on aesthetic experience. He claimed that objective viewing of a painting requires that "we need bring with us nothing from life, no knowledge of its ideas and affairs, no familiarity with its emotions." The circle around Picasso felt an urge to liberate themselves from the traditional observer's perspective in their representation of the world. They were strongly interested in non-Euclidean geometry. The result was Cubism where the artist tries to interpret the world by moving around and then capture several successive appearances of that world. Consequently a painting will reflect multiple viewpoints at the same time.

Miller also claims that traces of the same discussion on relativism appears in modern linguistic albeit not with the same strong emphasis. Two linguists Edward Sapir and Benjamin S. Whorf were early advocators of this view (cf. Whorf, 1956). Despite evident flaw in their fieldwork they gave rise to discussions on conceptual relativism that claims that different cultures with different languages have such different worldviews that they couldn't communicate with each other. This is a radical view that does not yield broad support. Theories of linguistic incommensurability are radical and must be considered a side movement. Yet there is clearly the notion of differences in point of view, but only with a kind of co-ordinate system that can act as reference across time. Causal theory permits meaning to change while the reference remains fixed. That is to say that the use of the language may change, but the ontology remains the same.

The above paragraphs illustrate that it is difficult to come to terms with the task of representing the world. Yet the emergence of relativism has come forward. Today we cannot escape the fact that Einstein was right. We are also intrigued by the Picassonian expression. Every person that has seen the picture “La Guernica” (see Figure 2) depicting the terror of bombardment on the Bask town during the Spanish civil war is possibly struck by the drama and the emotional aspect that the picture present. A close-up view of the various elements of the picture reveals a simple Cubist’ syntax – to some – a far-fetched abstraction very remote from the common understanding of a piece of art. Yet, step backward a few meters and the meaning and the emotion that it stirs emerges through the wholeness of the artistic expression. Move sideways and you will catch a glimpse of a perspective that was not immediately apparent from the first point of observation. Move back and forth in front of the picture and your movement will impose a kind of animated impression of an unfolding drama. This is the genius of Picasso. He makes the audience participants. The experience that we attribute to Picasso is nevertheless a personal one.

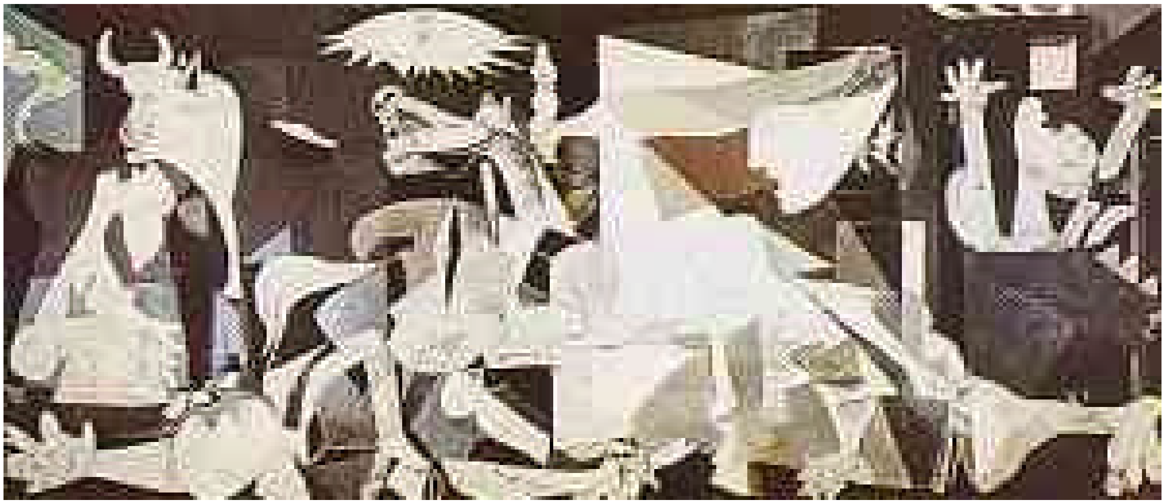


Figure 2: “La Guernica”, Pablo Picasso, 1937.

In their seminal work on knowledge building (Nonaka and Takeuchi, 1995) regard the traditional scientific approach of the Western world as an obstacle to understanding the tacit nature of people’s knowledge and how it can be shared. They throw a critical light on the Westerly tradition to objectify nature. They point out that there is a major difference between Westerly philosophy and Japanese philosophical tradition, that could be the major reason for a different understanding of what knowledge types should play a major role. Japanese epistomology has nurtured a delicate and sophisticated sensitivity to nature, it has prevented the objectification of nature and the development of what they call “the Western skepticism”. Despite the fact that it is hard to look past a currently common reasoning paradigm, even scientists in the West could support such a claim. On top of that, (Briggs and Peat, 1999) call for a new perception of the world that looks at wholeness instead of considering a separate “objective” truth as standing next to various “subjective” perceptions of the world. Briggs and Peat cite the writer and physicist Fritjof Capra who claims that the human race is experiencing a “crisis of perception”. This perception yields a fragmented, analytic view of reality that is inadequate for dealing with our overpopulated interconnected world. He attributes this crisis to the traditional mechanical view of the world in the Western society so strongly advocated by Descartes.

Nonaka and Takeuchi are quite clear about the influence that world perception has on language. Basic attitudes associated with the "oneness of humanity and nature" in Japanese epistemology can also be found in the structure of the Japanese language. Physical and concrete images of objects are indispensable for Japanese expression. Japanese think visually and manipulate tangible images. In Japanese statements made by the speaker articulate certain concrete images. The Japanese language is characterised by visual concepts that are highly context-specific in terms of both time and space."

To draw this line of reasoning somewhat further: Zen Buddhism formulated the principle of oneness of body and mind. According to Nonaka and Takeuchi, Zen profoundly affected the Samurai tradition. First principle in Samurai education was being a man of action. Cognising the world in terms of words is not central in this samurai and Zen philosophy, something that Nonaka claims has coloured the general thinking in Japanese society today.

To speak with (Nonaka and Takeuchi, 1995): "The inherent characteristics of the Japanese language reveal a unique view of time and space. The Japanese see time as a continuous flow of permanently updated "present". Many Japanese novels do not have any fixed time point in their plots, and traditional Japanese poems are free from any fixed time perspective. In contrast, Westerners have a sequential view of time and grasp the present and forecast the future in a historical retrospection of the past. The Japanese view of time is more circular and momentalistic. Everything appears and disappears occasionally and ultimate reality is confined to here and now. The Japanese view of space is also free from a fixed perspective, as is clearly depicted in traditional Japanese art. Since the perspective is not fixed there is no need to draw shadows.

A typical Western way of perceiving the world is to conceptualise things from an objective vantage point. Opposed to that, Japanese would rather relate themselves to the things and persons that the world encompasses. Hence there is an interpersonal acceptance of the world as it is. A message is often transferred through the use of context, not solely by a self-complete grammatical code. This can be seen from the fact that verbs in the Japanese language do not conjugate with the subject of the sentence. In Indo-European languages, verbs basically conjugate in accordance with the subject because the meaning of a verb are always used in the same form in any context. The perspective of the speaker can thus be shared naturally and smoothly by the group because of the sympathetic nature of the verb."

1.3 Implications on the present work

Truly traditional Japanese thinking clearly converges with ideas that have gradually emerged in certain camps in the West since Frege. However, this also tells us that as member of the Western tradition we may carry ballast that makes it difficult to approach the NLP problem in a manner that reveals a solution. Early work focused very much on the structure of sentences. During the 70's and 80's context-free grammars were the order of the day. "Divide and conquer" was the motto, and many such strategies were advocated. Studying single words was seen as most appropriate. Despite lesser corpora and costly and weaker computing power, early more or less fruitless attempts to NLP seem to stem from weakness in methodology rather than limitations in experimental facilities.

For reasons discussed above, it is felt to be necessary to shift position towards a more holistic, non-linear approach that accepts the limitations as well as the strengths of the Indo-European languages. Inspired by the fact that other cultures place different emphasis on grammatical structure and the role of temporal and structural roles and still communicate we have come to view a language such as English as a two-dimensional reference (a stream) to a subjective perception of the world. The underlying principle is that the world we live in, interact with and try to understand is dynamic and highly multi-dimensional in its appearances. This is clearly reflected in natural language by the use of prepositional expressions as the common way of pinning down an event or thing in terms of time and/or place. The ontology that we are aiming for is the world, but we perceive that the reference to this can never be objective. The way we should proceed in order to capture the essence of experience and knowledge from written documentation resides not so much with the absolute definition of the words applied, but with their usage instead. Hence the importance of the taxonomic aspect of the representation that we are seeking become less dominant. However, more important are the lateral associations between the words applied.

1.4 Schools of Linguistics and cognitive psychology

1.4.1 A retrospective look

In the discourse of natural language processing it is essential to round up the basic positions within the fields of linguistics, computational linguistics, cognitive psychology as well as AI. There is a strong interrelationship between the different sciences. They all share in different ways views on intelligence, communication, language structures and intelligence. In various ways they have influenced the whole domain of information retrieval, information extraction and text analysis.

The field of linguistics has long traditions. The earliest reference we have found beyond philosophy with relevancy to our own discourse is the work of the Norwegian linguist Ivar Aasen (Aasen 1848) who created an artificial grammar in the 19th century based on multiple native dialects. The task undertaken aimed at creating a written language based on a synthesis of spoken dialects. The result came to be known as New Norwegian[1] and is one of two official languages in Norway today. Aasen's work set a standard for similar attempts and his grammar constitutes the basis for a multitude of research that seeks to define a written language for minorities and new nations in the third world. What is interesting to note here is that research of this kind is motivated by the need to create a national identity, a common means of communication incorporating native tongues of different kinds. This endeavour starts by building word constructions from phonemes and developing a grammar for linking these together for different modes of speech. Through the 19th century linguistics were closely tied with philosophy. The work of Frege (Bremdal 99) sees a close bridging of this.

During the latter part of the 19th century linguistics become more experimental. This is closely related with the separation of psychology from philosophy. Theory was not enough. At the turn of the century psychology established itself as an independent science through use of novel methods and new ideas. The gradual refusal of introspection as a scientific method contributed to the divide. Like for so many other sciences empirical studies emerged as the backbone of the research undertaken. There goes a thread from the ideas that we are treating here back to the work in psychology presented by James (James 1890) and Thorndike (Thorndike 1898) in the 1890ies. Psychology took on different guises that turned into rival schools. The manifestation of psychology as an independent science came with the rise of a field of psychology termed "Behaviorism" forefronted by Watson in 1913. (Watson 1913) A major focus of linguists was related to children's learning of their first language. Several experimental studies were launched to improve our understanding of the acquisition of a language. Hence a part of linguistics and psychology became united in method and idea.

In order to understand the basic paradigms of computational linguistics, information retrieval and extraction it is important to understand the influence from psychology not only in terms of

[1] as opposed to the written language imposed by 400 years of Danish rule

methods and ideas, but also in terms of the rivalry within psychology itself. Behaviourism represented a main stream activity in the first half of 20th century. Its early proponents called for a more empirical practice that could treat important subjects in a more rational manner than philosophical theorists. They discarded the concept of mental activity and focused on basic perception in what has become known as stimulus-response systems. Behaviourism can be seen as an extension of an early field of psychology called accosiationism. In addition to this there was the Freudian school and Gestalt psychology. Through the years we see the emergence of mathematical psychology and cognitive psychology. The years after the Second World War sees what Newell has called the "cognitive revolution" (Newell, 1980). Before this all science related with human psychology and first language learning had been dominated by behaviourist influences. Since Watson experimental psychology had focused at learning through exposure to various kinds of stimuli or impressions. Important landmarks were made in understanding mechanisms for adaptation and response. All of these had in common the fundamental concept that organisms exposed to a stimuli acted upon such a stimuli with no involvement of any intermediate processing prior to the response process itself. In linguistics statistical methods emerged as a common means for building various types of grammars. This had emerged early in associationist psychology (Carr 1931) In first language learning principal ideas circulated around the issue of environmental impact on word acquisition. Researcher like Zipf (Zipf, 1935), McCarthy (McCarty, 1952)[2], Harris (Harris 1961) had in turn profound impact on this school. Despite its pronounced position during the 1920's and 1930's behaviourist ideas were not left unchallenged. The German Gestalt psychologists like Kohler and Wertheimer launched heavy critics on what they tended to address as a "limited view" on perception. A more silent opposition was formulated by Duncker in his work "*Zur Psychologie des produktiven Denkens*" (Duncker, 1935). Zipf also received a barrage of critics from different positions for his early thesis work on linguistics. Another important piece of work that came to influence empirical work was what later became established as information theory. Claude Shannon published in 1948 (Shannon, 1948) his mathematical theory of communication where he used probabilistic models to describe his noisy channel theory. Shannon's work showed that probabilistic methods could be applied to compress a sequence of data to a certain minimum given by the entropy of the input and the channel width. He also showed that this noise channel model could be used to estimate the likelihood of a given set of characters in a sequence. This work has had a deep effect on empirical work in linguistics, but received little attention beyond the behaviourist camp. Rather at the end of the 1950's the field of linguistics and psychology would see a paradigm shift with the cognitive revolution. Already at the time ideas that were already established within the framework of Gestalt psychology had been proposed in the early work on management theory and human information processing by Herbert Simon (1972). Lashley (1960) attacked the behaviourists and argued that human behaviour demonstrates a recursiveness that cannot be explained without the concept of mental control. The work of Miller (1956) and Lashley had laid the foundation for a theory of cognition. The cognitive view saw man as an information-processing organism that could reflect and manipulate information symbols in the mind and act according to inferences made upon these. When Skinner (Skinner, 1957) released his work "Verbal Behaviour" it triggered a reaction. Skinner tried to explain how children learn to speak based on behaviouristic principles of learning. In direct response to Skinner's work Chomsky published his critics (Chomsky, 1957) that paved the way for closer attention to cognition as a fundamental aspect in human psychology. He introduced his competence approximation. Chomsky was sceptical that approximate methods such as Shannon's n-gram approximation where inappropriate for his work on a transformational grammar. Chomsky's critics eroded the research fundament of behaviourist inclined research like that of Harris. Based on mathematical theory of automation Chomsky showed that Skinner's principles could only be established in terms of a finite state machine, that is, only with a finite set of symbols, instructions and memory states. He also implied the productive nature of language. There is no limit to how many and how long sentences we can construct and interpret. Chomsky introduced a transformational and generative grammar that if fully developed would specify every legal language construct possible.

The new ideas brought forward by linguistics such as Chomsky and the cognitive psychologists encouraged other researchers to take advantage of the experimental power of the early digital computer to test out their early ideas. This in turn gave rise to a new science and engineering discipline called Artificial Intelligence headed by leading researchers like Marvin Minsky, Herb Simon and Allen Newell. The cognitive or rationalist tradition has dominated the area of psychology since the sixties, but loosing significant terrain in later years.

In the 1990's work on empirical approaches in linguistics have accelerated once more (Armstrong-Warwick 1994, Church 1994). The type of work that was very common in the 1950's and earlier have been revitalised. There are several reasons for that. According to Church the single main reason is the success of simple statistical based methods in speech recognition. Several attempts to apply cognitive models and ATN type of grammars (see below) failed. Work on speech recognition that IBM initiated in 1972 failed to produce the expected results. However, by capitalising on Shannon's noisy channel model and work like that of Harris significant progress was made. This spurred a broad and renewed interest in empirical methods for use in computational linguistics at large. Church also points to the tremendous increase in available computing power and large on line corpora such as the Brown Corpus (Francis and Kucera 1982) as contributing factors to the renewed interest. Yet no school can claim to be even close to the ultimate solution.

At the turn of the century we see that different types of work are coming together. The division is still there, but failures and successes of the past in both camps have evened out the differences. There is clear evidence that the truth in terms of natural language understanding cannot be monopolised by any of the communities. Time and hard work have revealed basic flaws in both theory and practice regardless of the fundamental inclination. We are still far off from the perfect machine translation system or the computer understanding. The state-of- the-art is greatly influenced by increased and affordable computing power and the emergence of the Internet. In the following we will look at the important principles behind different initiatives. We will use this as a basis for a discussion on current developments and the fundamentals of the state-of-the-art in terms of information extraction and ontology building.

1.4.2 Different models of perception

In this discourse it is important to understand the different views on human perception as seen from a purely behaviourist view and the traditional cognitive view. Figure 3 depicts the "Behaviourist View" in a simplistic way. The basic conception is that the environment provokes a response from the subject through repetitive exposures. Adequate responses are encouraged through a reward policy or a credit assignment principle that the lies inherent in the environment itself.

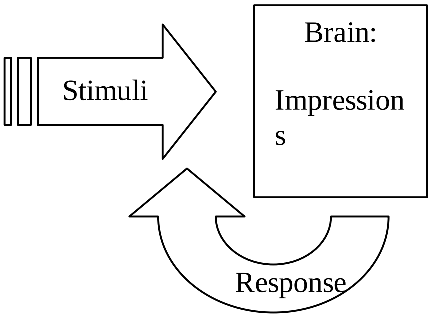


Figure 3 The Behaviourist View

The Behaviourist sees the brain as a basic storage medium for patterns or impressions that yields negative or positive stimuli. Addressed in terms of linguistics the model appreciates the relationship between parents' use of words and children's ability to learn and reapply the same words in order to take part in communication (Wolff 1991)

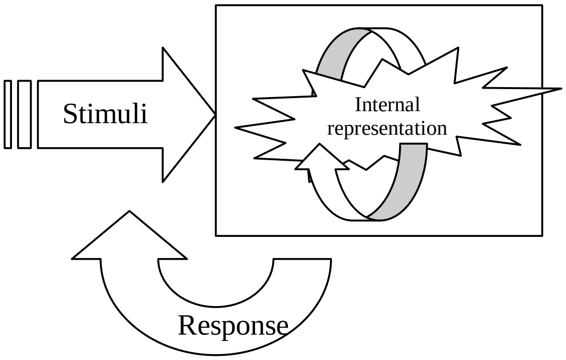


Figure 4 The Cognitive View

The cognitive view recognises the mind as an information processing unit and actually stores a representation of the environment that acts upon it. Both schools recognise that information must be perceived and some action performed in order to close the cycle. Yet the cognitive view emphasises an intermediate step that grants the brain a far more important role than that of the behaviourists. The notion of mental process is kept very strong along with the widely held belief that discrete symbols represents the fundament for internal reflection and processing, and thus understanding.

In spite of their simplicity the figures above convey the essential differences. These differences have coloured research in both linguistics, psychology and artificial intelligence for the full past century and the differences are still unsettled. Despite tremendous activity during the past thirty years cognitive psychology has not emerged as a single triumphant community. On the contrary neo-connectionism and the success of statistical natural language processing (Manning, 1999) calls for a revitalisation of methods that emerged as long as 70 years ago.

1.4.3 Natural language processing and artificial intelligence

Already from the start natural language processing became a very important area of research within the emerging science of artificial intelligence. For the better part of the thirty years between 1960 and 1990 this effort was dominated by ideas stemming from cognitive psychology and Chomsky. But it is important to point out that the use of computers in terms of linguistics was not new. The success of allied cryptographics during the second world was and the accomplishment of von Neumann stirred broad optimism in natural language processing and translation by means of computers. However, around 1965 a feeling of chimera was spreading. Terry Winograd recalls:

"In the mid 1960's natural language research with computers proceeded in the wake of widespread disillusionment caused by failure of the highly touted and heavily funded machine translation projects."

At that point computer related work had been preoccupied with syntactic word shuffling. Most endeavours focused on analysis of syntactic structures and identification of lexical terms. This was clearly not sufficient. Programs had to deal with what the words and the sentences meant.

As mentioned earlier Chomsky criticised the empirical efforts in main stream psychology with a strong emphasis on language learning. A major part of his reproach was directed towards the taxonomic methods based on statistics. His own efforts had been focused on a transformational grammar that was strongly rooted in the cognitive model. His position can be clearly reviewed from this statement:

"There is surely no reason today for taking seriously a position that attributes a complex human achievement entirely to months (or at most years) of experience, rather than two millions of years of evolution or to principles of neural organisation that may be even more deeply grounded in physical law. " (Chomsky, 1965)

The predominant idea expressed here is that language learning must be seen as an evolutionary consequence. It cannot be learned from scratch. The language ability is hard wired in a genetically form. This is the fundamental aspect of a "Universal Grammar" which seeks to identify the triggers that mobilises the inherent language capabilities that each person carry with him. Chomsky's criticism of the statistical approaches and his work on the transformational grammar had a profound effect on work in several related camps. This also included research on natural language processing within the AI laboratories. AI inherited the context-free parser techniques and the Chomsky's theories. But AI was focused primarily at computational intelligence. Knowledge and knowledge representation lied at its heart. It seemed obvious that by adding knowledge to the language parsing mechanism better results could be achieved. Work was initiated both at MIT and Carnegie-Mellon University to pursue this goal.

Two efforts stand out from this period, Dan Bobrow's STUDENT (Minsky, 1966) and Ross Quillan's Semantic Memory Program (Quillan, 1968). Both efforts produced technically interesting programs that could successfully handle communication within restricted domains. They indeed showed that adding a little knowledge would yield semantic, albeit limited, capabilities and improve output. However, their capabilities were grossly overrated by people like Marvin Minsky and used as ikons for a new era leading to high fidelity computer understanding. Such claims stirred reactions from people like Hubert Dreyfus (Dreufus, 1979) and later boomeranged on this type of research in a very negative way. One reason beyond general inflation of performance was the fact that these early systems still operated basically on syntax.

Winograd's work in 1971 (Winograd, 1972) is often used as a prime reference in terms of AI research. The doyens of AI like Marvin Minsky as "a major advance" heralded it. Winograd himself returned to rate his own achievements nearly 20 years later (Winograd and Flores 1986).

“SHRDLU is based on the assumption that the meanings of words and of the sentences and phrases made up of them can be characterised independently of the interpretation given by individuals in a situation. There are, of course, aspects of meaning that call for qualifying this assumption. A sentence may include indexicals (words like I you now) whose referents are elements of the conversation situation, or may have connotative effects that depend on the full understanding and empathy of the hearer. But these are seen as add-ons to a central core of meaning that is context independent.“ He continues: “ The concept of literal meaning is inadequate even when dealing with the most mundane examples. Meaning always derives from an interpretation that is rooted in a situation.” In this manner he accords with Hubert Dreyfus, one of his principal critics from the early post-SHRDLU days.

Parallel to the work at MIT (Minsky) and Carnegie Mellon University (Simon), a new research community surrounding Roger Schank at Yale emerged. Schank’s work was deeply rooted in cognitive psychology, but with a slightly different basis than his peers at MIT and CMU. His basic focus was memory structures. In order to tackle the issue of semantics and apply it in terms of natural language processing it was important to build memory structures that could embrace episodic experience. This addresses fundamental issues of knowledge representation. In his Conceptual Dependency Theory (1975, 1977) Schank attempted to define a plausible representation that could support language use. He wanted a representation that was unambiguous and unique. His aim was to express the meaning of any sentence in any language. The representations were intended to be language independent. Two sentences with the same meaning should have the same representation whether they were paraphrases within a given language or translations between languages.

In this effort he defined a set primitive acts from where all aspects of speech could be derived (see Table 1). In addition he introduced the script as the basic representation. A script would capture the essential concepts associated with an episode. The notion of a script was later revised to become a non-static feature composed of lesser structures that Schank called MOPs (Memory Organisation Packages, Schank 1982). We should note that several of these features were actually applied to build the first commercial NP application, ATRANS, which handled bank transfer statements in free text. The early ideas of Schank spawned during the 70’ies and eighties becoming very influential in terms of AI based natural language processing (see Riesbeck 1986, Wilensky 1983, DeJong 1979 , Lebowitz 1983, Lange 1999).

1.4.3.1 Five physical actions of people

PROPEL (apply force to)
MOVE (move a body part)
INGEST (take something inside of an animate object)
EXPEL (force something which is inside an animate object out)
GRASP (grasp an object physically).

1.4.3.2 State changes

PTRANS (change the location of something)
ATRANS (change some abstract relationship i.e. ownership)

1.4.3.3 Instrumental acts

SPEAK (produce a sound)
ATTEND (direct some sensor towards a stimuli)

1.4.3.4 Mental acts

MTRANS (transfer information from a source to a receiver)
MBUILD (create and combine thoughts)

Table 1 Schank's primitive acts

The ideas that can be first of all attributed to Schank’s dynamic memory theory is a knowledge rich, memory modelling approach that has worked well for particular applications such as the news interpreting systems.

A basic issue associated with all knowledge rich methods based on the cognitive concept is representation. The production rules of the GPS (Newell and Simon 1972), predicate logic (Kowalski 1979), the frame structures of Minsky (Minsky 1975) are like Schank’s script types of representations that can be applied to build the model that holds the knowledge required for the interpretation job. Yet, the field of knowledge representation is not a settled issue. Both the ontological aspect and the epistemological issue remain much debated. As shown in Figure 3 and Figure 4 the representation constitutes a unique feature with the cognitive model and thus main stream AI. However, Brooks (Brooks 91) argues that this issue does not have to dominate if intelligence is approached in a step by step manner, focusing strictly on the interface between the intelligent system and the rest of the world. This interface is demonstrated through coherent combinations of perception and action. Here we see that Brooks stretches out a hand to the behaviourists. His work also suggests that the memory does not necessarily have to be confined to the mind. Low-level organic intelligence as well as humans apply external memories to perceive and to respond. Another important point drawn from Schank’s Dynamic Memory theory is that the memory structures changes with new experience. If the same memory structures are mobilised to process language we could be convinced that the effort is very much based on the individual and the way he represents his experience. The language-processing task can no longer be looked upon as an objective effort.

1.5 Basic linguistic theory

A comprehensive introduction to modern linguistic theory as advocated by Noam Chomsky can be found in Haegeman (Haegeman 1991). A summarised version of this is presented in Winston (Winston 1992).

The basic concept in this discourse is the constraint theory related to how words and sentences fit together. Linguistics aim to embrace a full theory. In this pursuit it is imperative that they look at sentences with a representative verbal structure. They tend to avoid simple sentences because explanations for such would be too specialised. At the heart of this traditional theory lies the X-bar schema (see Figure 5). The X-bar schema defines an analytic method for decomposing sentences into phrases and words. It also defines a representation that exposes grammatical constraints in a unified and simple manner. Many linguistics maintain their faith in the X-bar hypothesis. They believe that it offers the right perspective for determining what is in the "Universal Grammar" because it provides a fair way of expressing constraints.

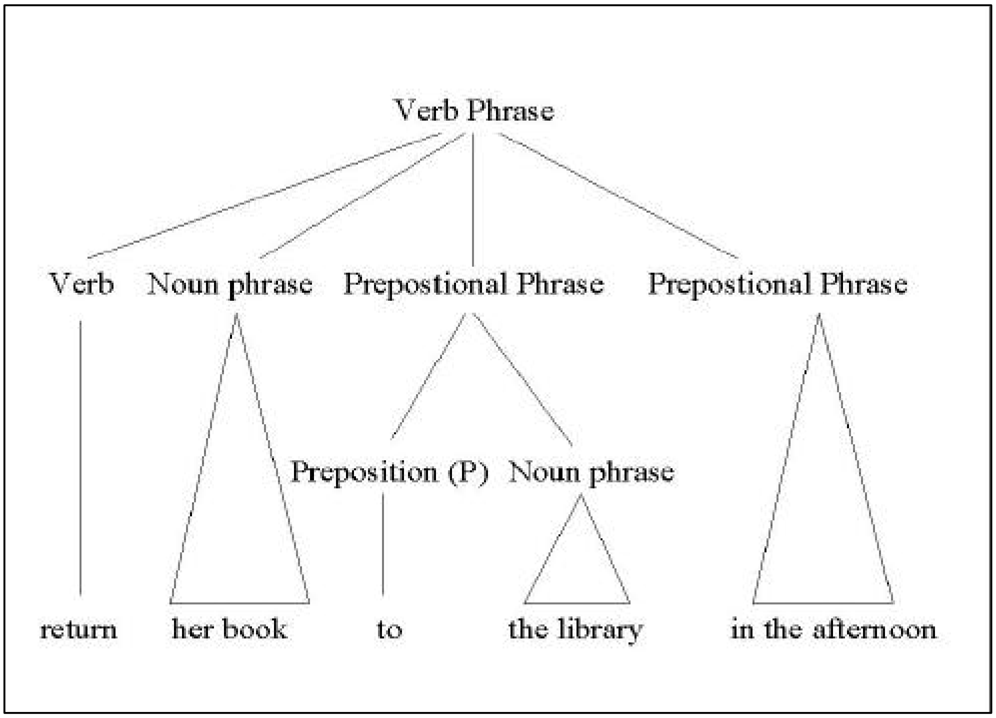


Figure 5 ***Instance of X-bar schema*** *for "return her book to the library in the afternoon" (after Winston 92). The sentence and phrases are decomposed into grammatical primitives. Word groups whose internal structure is not shown are indicated by triangles*

The decomposition agenda seeks to identify the basic grammatical primitives and word groups that include these primitives. Primitives include verbs and nouns. Word groups would be verb phrases, noun phrases and prepositional phrases. In Figure 5 we have shown a classic instance of the X-tree concept for the phrase "return her book to the library in the afternoon". Improvements on this are achieved by turning this into a binary tree structure (see Figure 6). A binary tree structure will simply constrain tree branching to two. Hence the verb phrase above would be represented as a trunk while each lesser word or word group would branch out from the trunk one at the time. The leaves on the tree are single words. This binary-tree hypothesis can be applied at all levels. It takes advantage of the fact that many phrase types have the same structure. It simply enables an analyst to spot replication possibilities and commonalties in the sentence structure. The idea is that a subset of a sentence can be replaced by another subset if they observe the same grammatical constraints exposed by the binary-tree. It also exposes extension possibilities at a given phrase level. One useful rule is the fact that words brought in from the right constitute complements to, for example, a noun phrase, while a word brought in form the left is called a specifier[3].

A fundamental hypothesis associated with the X-bar schema discussed above is that "All Phrases

[3] In Russian and Japanese this rule will not necessarily hold though it can be observed in both in the Scandinavian languages as well as in German.

Have the Same Structure". This has profound effects on the aspect of inflection were verb forms can be replaced with no concern to the other structures in the sentence while changing the expression in terms of tense information. One example is the sentence "Sarah will return her book" and the variant "Sarah returned her book". Inflectional phrases like this display the same structure in the binary X-bar tree. Another concept that linguistics of this school apply is the complementiser phrase (CP) which is headed by the word "that". Inclusion of an inflection phrase like one of the above in a complementiser phrase makes it convenient to make transition of phrases across the tree. A simple movement of the word "will" from the inflection phrase's head position to a landing site in the complementizer phrase's head position enable the question "Will Sarah return her book?" (see Figure 6). This explains both the relationships between a statement and the associated question as well as to what liberty other word groups can be moved around without violating the basic constraints.

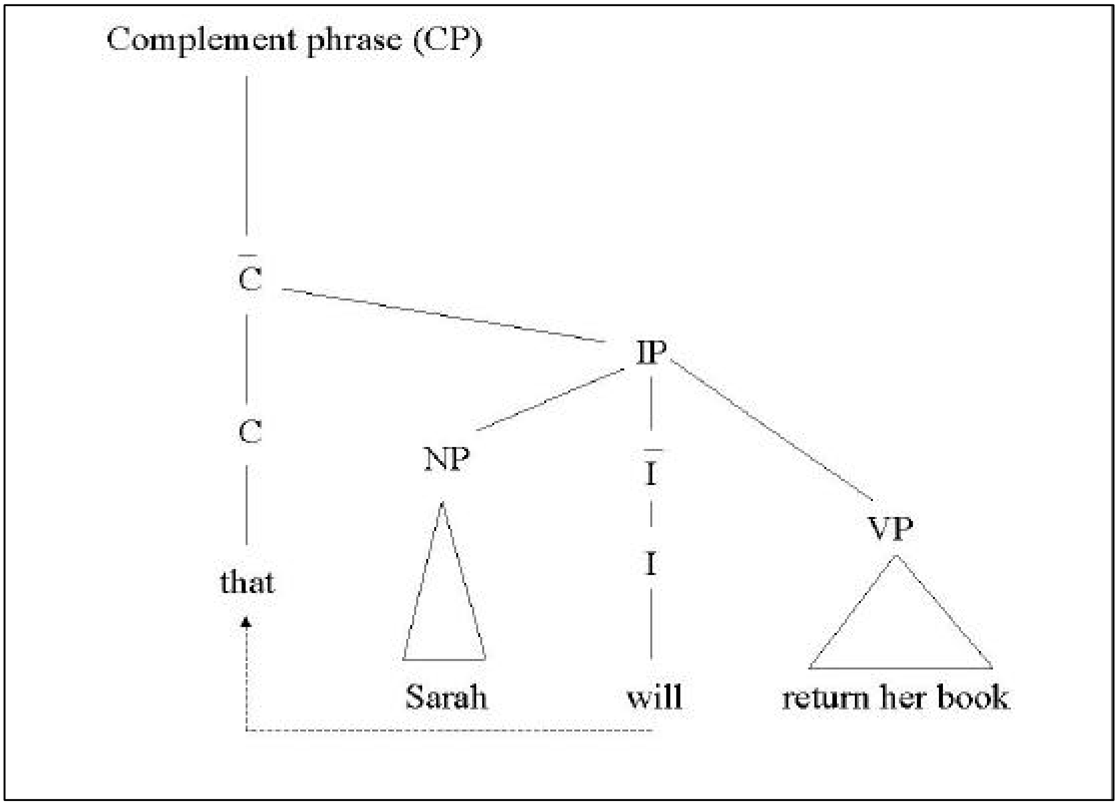


Figure 6 A binary X-bar tree instance showing the convenience of a complementizer. By moving the inflection from the IP to the CP a question can be formulated on the same subject.

The concept of "governing head" or "government" is another property determined by the X-bar tree. The governing head is the node in the tree that governs the case of the phrase. Hence it determines how, for instance, pronouns can be substituted for each other. The X-bar theory also handles several other important constraints. These include the handling of "wh-words" i.e. when, where, what in terms of the subjacency principle which is a constraint that specify how questions can be constructed with these words.

The X-bar theory holds a strong position in the linguistics community and it continues to inspire AI researchers. However the theory focuses entirely on language competence. That is, its mere focus are the constraints that define how words can be put together. This far off from the focus of Schank and similar initiatives within the AI camp who are trying to build models of language performance, programs that in some sense understands the content expressed in some language. Even within the study of language competence there is a lot of debate on different points of view.

1.6 Context-free grammars

Context-free grammars such as the ATN grammars (Augmented Transition Net Grammars) have traditionally been popular with computer people (see Figure 7). However, linguistics hold the opinion that they are poor representations of linguistic knowledge. They are usually sufficient, however, to represent the type of linguistic knowledge that children learn at school. The ATN concept was introduced in the machine intelligence community in 1968 by Thorne et al. It was later made popular by several other pioneers within computer based natural language processing such as Bobrow and Fraser (1969). ATN's have been applied with some success within a constrained area of application. The ATN consists of a number of nodes linked together by a net. Each link corresponds to a phrase or a single word. The various nets can be nested at the nodes. Some links can be made mandatory while others are optional. An ATN is basically a syntactic instrument that incorporates basic linguistic knowledge about English. However, the ATN can also be used to handle basic language performance tasks. Winston addresses an example where a syntactic transition net is applied in order to construct English sentences that drive database retrieval. This and other examples demonstrate how a move to a semantically oriented transition system picks out the content elements from a sentence and constructs a query that retrieve the information from a database. Although dashing the complexities of the English languages such ATN's represent early attempts of information extraction for tangible engineering tasks.

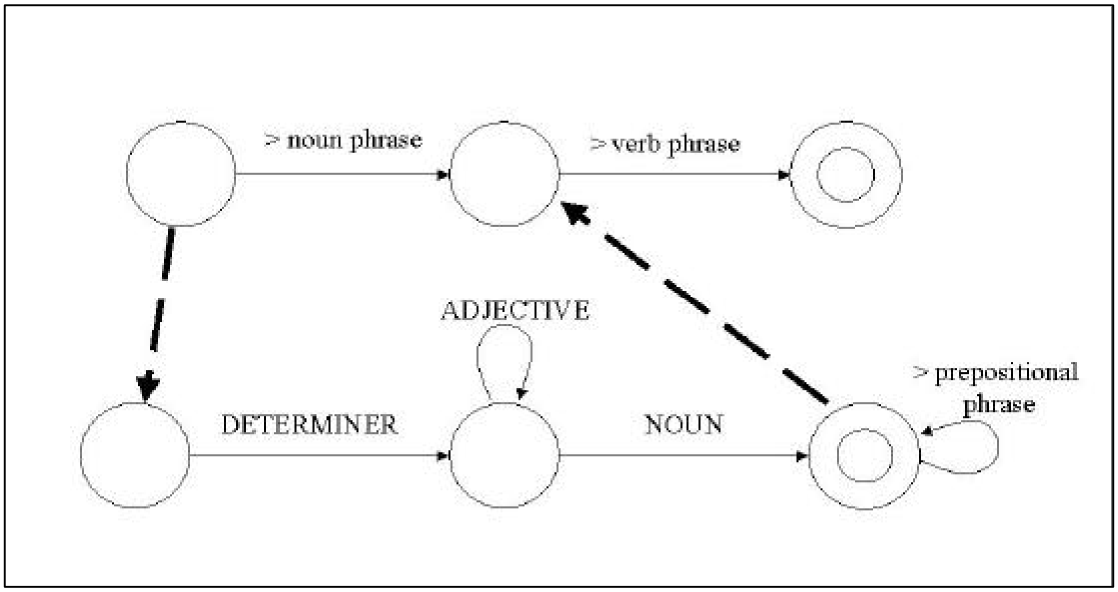


Figure 7 A simple ATN showing how the noun phrase link can be expanded

1.7 Bringing more knowledge into the analysis

As pointed out earlier the AI community looked to knowledge about the world and the issue of semantics in order to cope with the early failures of natural language processing. Several different approaches were undertaken. One important effort in mentioning is the work by Yorick A. Wilks who was probably the first to undertake computer related work on deeper level analysis of sentences (Wilks 72[4]. However we believe that the work of Schank and the work on scripts and MOPs stand out because of its impact on research on semantic analysis of language through the 70's , 80's and 90's and because of its focus on memory structuring that we believe is essential to handle both semantics and pragmatics of natural language performance . Several spin-offs from Schanks's early ideas could be mentioned including work by DeJong on the system FRUMP (1979), the Direct Memory Parsing by Riesbeck (1986), the RESEARCHER system of Lebowitz (1986) and memory based word disambiguation by Lytinen (1986) as well as later work of Lange and Wharton (Lange 1999) and Ashwin Ram (1999). Here we will illustrate the basic concepts with a simple run-through on work presented by Michael Lebowitz's (1983) seminal paper from 1983. In his IPP project Lebowitz introduces a version of Schank's MOP's that he calls S-MOPs (simple MOP). The IPP is a computer system that is designed to read and generalise from large number of news stories. Stories like these are typically centred around reports on particular events. Events are represented as S-MOPS and Action Units (AU's). Both AUs and S-MOPs are frame-like language-free conceptual structures. In this respect they serve the same roles as the semantic primitives in the Conceptual Dependency framework introduced above. Yet AUs and S-MOPs are strongly domain-dependent and idiosyncratic. According to Lebowitz it is this specialisation that provides the power in processing. Both concepts are applied for so-called memory-based predictions. Given an AU and S-MOP the IPP system had to identify tangible objects that could fill in the roles defined by the AU or the S-MOP. This could be people, groups and places, concepts that Lebowitz call "Picture Producers". Lebowitz points out:

"In any particular domain, there are a relatively small number of memory structures (the S-MOPs and Action Units in IPP). In contrast, there are thousands of lexical items and an unlimited variety of syntactic constructions that can indicate the relevance of those memory structures. Thus, it is certainly desirable to make as much processing as possibly relatively independent of the text being read. This allows most of the necessary information for understanding to be organised around the small number of memory structures instead of the many lexical items. The parsing process need only rely on minimal information from specific lexical items, using them to identify Action Units and Picture Procedures."

Below we will briefly present an example used to present the IPP system. The following story is given:

"S18 (UPI, 23 July 80, Lebanon)

Gunmen today shot and killed Riyad Taha, president of the Lebanese newspaper publishers' association and his driver in an ambush. Taha, 54, a Shiite Moslem, was shot as he was going to his office in predominantlyu Moslem West Beirut.The gunmen who opened fire at his car escaped."

The processing of the story is initiated with no specific expectations of its content. The first step is to define the context that will allow memory-based rules to be used. This can be done almost immediately with the word "Gunmen" as the first source for generating expectations. This is a token maker that triggers several associated Action Units. In order to select the proper AU it uses other lexical elements to help screen its prediction. Once the proper AU is found it is instantiated. "Gunmen" is assigned to the ACTOR role of the AU. When "shot" is read the prediction process associated with the selected AU yields an answer. An AU named "shoot" associated with gunmen is instantiated. Given this context new templates are identified in order to understand later text. Concurrently an S-MOP that encompasses what has been determined is searched for. Once this overall structure is defined the system will attempt to fill in the empty slots. Identifying the Riyad Taha of the first large noun phrase it triggers expectations which assume Taha to be a victim both for the "gunmen" and the "shoot" templates. The system continues to work on subsequent phrases trying to fill the slots with tangible references. Until the parser reaches the word "ambush" the system is not able really analyse much. But the word "ambush" triggers another AU which can be attached to the overall S-MOP. The former procedures are iterated yielding the final story representation:

```
Ev1 =
        MEM-NAME   S-ATTACK-PERSON
        ACTOR              GUNMEN
        VICTIM             54 YEAR-OLD RIYAD TAHA/PRESIDENT OF THE
                   LEBANESE NEWSPAPER/PUBLISHERS'
                   ASSOCIATION/SHIITE MOSLEM
METHODS
EV0 =
        MEM-NAME   $SHOOT
        ACTOR              GUNMEN
        VICTIM             54 YEAR-OLD RIYAD TAHA/PRESIDENT OF THE
                   LEBANESE NEWSPAPER/PUBLISHERS'
                   ASSOCIATION/SHIITE MOSLEM

EV3 =
        MEM-NAME   $AMBUSH
        ACTOR              GUNMEN
        VICTIM             54 YEAR-OLD RIYAD TAHA/PRESIDENT OF THE
                   LEBANESE NEWSPAPER

RESULTS
EV2 =
        MEM-NAME   $CAUSE-DEATH
        ACTOR              GUNMEN
        VICTIM             54 YEAR-OLD RIYAD TAHA/PRESIDENT OF THE
                   LEBANESE NEWSPAPER
        HEALTH     -10
EV4 =
        MEM-NAME   GS-ESCAPE-TERRORIST
        ACTOR              GUNMEN
```

Subsequent research based on this has since branched off in two main directions. One has tried to integrate the syntactic analysis into a memory driven semantic analysis like that described above (Lebowitz 1986). The other avenue is that of building more opportunistic (Ram, 1999) and less context specific indexing mechanisms. MOP like structures represent stereotypical sequences of events very well suited for narrow domains, but they do not scale up well for broader applications. But according to Domeshek (1999) the most successful systems in recent message understanding conference (MUC) competitions were all essentially MOP or script based. The MUCs favours a limited focus. The script and MOP based approach has also constituted the basis for intentional analysis. Work like that of Wilensky (1983) is an example of that. Another trend has been to combine the memory oriented efforts with connectionist models such as that of Langston (Langston 1999).

1.8 Statistical approaches

As already discussed previously statistical linguistics may be regarded as a part of the empirical movement rooted in behaviourist psychology. Through the last century the empiricists gathered intriguing evidence that frequency of words and word patterns play an important role in terms of language and communication. Work of Carr (1931) and Shannon (Shannon 49) support this claim. Zipf (1935) formulated his Principle of Least Effort which argues that people will act so as to minimise the work that is not required instantly. According to this both the speaker and the hearer are trying to save their effort. Having a small vocabulary of common words conserves the speaker's effort and having a large vocabulary of individually rarer words in order to avoid ambiguity lessens the hearer's effort. This principle was founded on the statistical distribution in language that he established. This has become known as Zipf's law*:*

There is a constant k such that $f * r = k$ *where f is the frequency of the word and r is the rank.*

Word rank implies the relative order of a word in a given vocabulary. The reciprocal relationship between frequency and rank defines, according to Zipf, an optimum compromise, sometimes referred to cognitive economy, between the competing needs of the speaker and the listener. Zipf also established that rare words in most languages tend to be longer than common words. The variety of rare words is greater than the variety of common words.

Wolff (Wolff 1991) has made a thorough look at convincing cases from first language learning that undoubtedly yield support for a statistical initiative. In passing we will highlight certain observations about children. Wolff states that child's vocabulary increases rapidly, but evens out at early adulthood. Like most behaviourists he also claims that discretisation of speech where a spoken language is divided into words and other discrete segments like phrases and sentences is a capacity that is learned through exposure to the environment, basically that of parent's speech. Although challenged by many it implies that the kind of statistical analysis which can lead to a recognition of discrete segments in speech may also enable a child to learn non-verbal concepts – the meaning which lie behind the words of language and also learn the association between words and meanings. Ordinary objects in the world i.e. tables, chairs, cars, are segments of the visual world comparable with words and other segments in streams of speech. When exposure to references combined with experiences of such objects are repeated a vocabulary is developed that has clear semantic implications. But meaning is thus clearly a matter of use as pointed out by Wittgenstein. Another important observation that has been made is now known as the Law of Cumulative Complexity. According to Wolff and several of the references that he uses certain implications of such observations can be made. Although semantic knowledge may develop earlier than syntactic knowledge (or make itself apparent to the observer at an earlier age) it seems that the learning of both kinds of knowledge is integrated in a subtle way. One kind of knowledge is not a prerequisite for the learning of the other. This is an important message because it tells us that neither pure empirical nor pure structural approaches i.e. as defined by Chomsky may yield a final answer.

Wolff also points to studies that show that young children tend to respond with references to episodes while older children tend to refer to part-of speech relationship or a meaning relationship such as long-short and give-take. He also argues that children at the age of three experiences a "naming explosion" where they focus on object names – labels for things, animals and people.

According to a study by McCarthy in 1954 (McCarthy 1954) function words such as "of" tend to appear later in children's speech than content words despite the fact that function words tend to be the most frequent in any language. Children also do a systematic comparison of linguistic structures in an attempt to find elements shared by more than one structure. Wolff refers to Ruth Weir's classic pre-sleep studies of group of words where a young subject where massaging word structures to compile impressions from earlier that day into an abstraction hierarchy. This suggests that children do work to organise word exposures into some kind of taxonomy based on their occurrences in daily speech. The other aspect is the provable robustness associated with human listening. Church (Church 1994) uses the examples of "Joe is a rider of novels" and "Joe is a writer of horses". Most listeners would assume they heard "Joe is a writer of novels" and "Joe is a rider of horses". This phenomenon is sometimes addressed as "Hear What I Mean". Listeners usually have little problem with the diversity and ambiguity of speech because they know what the speaker is likely to say. The basic principle underlying this is that word patterns drive expectations and not memory structures as explained in the example borrowed from Lebowitz in the previous paragraph.

According to Church the Raleigh System on speech recognition created by IBM in the 70's created a revival of empirical studies. Shannon's information theory resided at the heart of the success they experienced. Although focused towards communication along a noisy telephone line the Raleigh System showed that Shannon's theory was very useful for speech recognition. Later it was applied for ORC (Optical Character Recognition), spelling correction systems and information retrieval.

Shannon's basic idea is that a sequence of good text (I) goes into a channel and a sequence of contaminated text comes out at the other end (O).

1.8.1.1.1.1 I -> Noisy Channel -> O

Shannon established a procedure where the original text, I, could be established from the corrupted output, O by establishing the set of all possible input texts and then selecting the text Î with the highest likelihood.

$$\hat{\mathbf{I}} = \mathbf{max_i\{ \quad P(I_i|O) \}}$$

or

$$\hat{\mathbf{I}} = \mathbf{max_i\{ \quad P(I_i)* P(O|\, I_i) \}}$$

Where :

I_i is the possible input and max_i the max score selection function

$P(I_i|O)$ is the probability that the given input of i produces the output O experienced
$P(I_i)$ is the prior probability of the input of i based on the frequency of its occurrence in a large population of words i.e. a word in a large corpus.
$P(O|\, I_i)$ is the channel probability which determines the probability that O will be presented at the output side when the given input i is presented at the front.

Adapted to part of speech tagging the noisy channel model can be written as

$$\mathbf{\hat{}p = max_i\{ \quad P(p_i)* P(W|\, p_i) \}}$$

p_i represent here the sequence of part-of-speech and W a corrupted sequence of words. This is a very complex function since it represents a infinite n-gram definition. However, it is possible to produce a good trigram approximation:

$$\mathbf{P(p_{1,} p_{2,} p_3 \ldots\ldots., p_n) \sim=} \mathrm{P}^{\mathbf{N}} \mathbf{P(p_i | p_{i-2} , p_{i-1})}$$

which yields

$$\mathbf{P(W_{1,} W_{2,} W_3 \ldots\ldots., W_n |p_{2,} p_3 \ldots\ldots., p_n) \sim=} \mathrm{P}^{\mathbf{N}} \mathbf{P(W_i | p_i)}$$

Where each word depends only on its own part of speech. The lexical probabilities, $P(W_i| p_i)$ and the contextual probabilities $P(p_i| p_{i-2} , p_{i-1})$ are both estimated by computing the statistics of large bodies of text. Thus the first parameter can be regarded as a dictionary and the second as a type of grammar. Church claims that traditional methods have tended to ignore lexical preferences. He states that they are the single-most important source of constraint for part-of-speech tagging, and are sufficient by themselves to resolve 90% or more of the tags. He uses the example “I see a bird” to illustrate this. In the 1 million words Brown Corpus the word “I” appears as a pronoun 5131 times out of 5132. “See appears as a verb 771 times out of 772. A appears as an article in 22938 times out of 22944 and bird appears as a noun in 25 times out of 25. This contrasts many of the ordinary dictionaries, which also lists a number of extremely rare alternatives.

In a general sense words in a corpus could be estimated according a binominal distribution which basically estimates the probability of a word’s occurrence versus the chance of no occurrence. Each word sample would represent and independent event similar to tossing a coin. Content words that typically constitute the backbone of a domain specific ontology will not follow this pattern. Content words tend to appear in chunks. Given one instance of a content word the likelihood that it will appear again in subsequent sentences is much higher.

Two important concepts in Shannon’s information theory and in statistical work on text are entropy, H, and Redundancy, R. Redundancy is a function of the entropy such as 1 – H = R. The entropy can be expressed as:

$$\mathbf{H = -K} * \mathrm{S}^{\mathbf{n}} \mathbf{P_i * logP_i}$$

Where
K is a constant
P_i the probability of the word i
Log is the logarithm to the base of 2 which yields the entropy in terms of bits.

The entropy defines the necessary bandwidth in order to account for possible combination of words. Said in a different way it characterises the degree of compression that is possible or to what degree input is redundant.

We can also express the cross entropy of some output versus input. This is illustrated here in terms of source code and the compiled version of a computer program:

$$\mathbf{H(source,code) =} \mathrm{S}_{\mathbf{s}} \mathrm{S}_{\mathbf{h}} \mathbf{P(s,h | source)* logP(s | h , code)}$$

As Church points out “P(s,h | source) is the joint probability of a symbol s following a history h given the source. P(s | h , code) is the conditional probability of s given the history (context) h and the code. The cross entropy defines the capacity of a language model to predict a source of data. If the language model is very good at predicting the future output of the source, then the cross entropy will be small.

The cross entropy thus defines a transfer function between some input compared to some output and this defines in a sense how the output responds to the input. It should be obvious that it is tempting to use this for parsing, language-to-language comparison (translation) and text-to-text comparison. Besides this the mathematical expressions described above can explain why crossword puzzles are tough while still being solvable. The latter is attributed to the entropy and the latter to the degree of redundancy in the language.

Information extraction and information retrieval can benefit both from the various observations on children as well as the principles laid down in Shannon's theory. The former suggests that the environment imposes a structural relationship between words that can be learned and in turn measured and estimated by means of mathematics. The principles of redundancy and entropy provide a basis for building performance-based grammars. Part-of-speech tagging can thus be achieved by statistical techniques such as Markov models. Manning (1999) shows how statistical techniques used on a large corpus of texts can expose principles exposed in the traditional X-bar schema. Despite Chomsky's verbal blitz on the inadequacies of Markov models in the 50-ies large on-line corpuses and powerful computers have proven their usefulness. But according to Manning Chomsky's criticism still holds. Markov chins cannot fully model natural language. One reason is their inability to model many recursive structures in a language. The process of information extraction in the context of statistics can be treated similar to grammatical tagging, but where the tags are semantic categories, not tags like *noun* or *verb*. However, a pure mechanistic approach cannot be pursued. " United States of America " must be treated as one construct, not four, to preserve the intended meaning. To identify and maintain collocations is therefore important. It is therefore interesting to note that it has always received significant attention by empiricists, yet is almost entirely neglected in structural linguistics. Collocations can be established through tagging or direct estimation using a n-gram type of formula.

For further illustration we can also point to how two techniques that can be used to infer meaning of sentences based on the principles discussed above. One is called *selection preferences* and another is called *semantic similarity*. Selection preference implies that some nouns are more likely to prefer some types of verbs rather than others. Manning (1999) shows an estimate for the occurrence of the noun "food" with respect to the verbs "eat", "see" and "find". This is given:

P(c) = P(food) = 0.25

For P(c|v):
P(food | eat) = 0.97
P(food | see) = 0.25
P(food | find) = 0.33

The conditional property of the verb "eat" with respect to other nouns in the corpus such as" people", "furniture" and "action" is 0.01, 0.01 and 0.01 respectively.

This yields a Selection Preference Strength for the verb "eat" with respect to "food":

SPS = 1.76

According to the formula:

$$\mathbf{SPS = S(v) = D(P(c|v) \,||\, P(c)) = S_c\, P(c|v)\log(P(c|v)(P(c))^{-1})}$$

(log to the base 2)

We see how this takes advantage of Shannon's theory on entropy. Since there is a clear tilt towards “food” and “eat” rather than food and anything else the entropy is low.

Semantic similarity is a type of generalisation, which consider the closest neighbours in generalising to the word of interest. The whole approach is based on the assumption that semantically similar words behave similarly. One example is how well the word “cosmonaut” can be replaced by the word “astronaut”. Another illustration is the use of the word “vehicle” instead of the word “car”. In order to determine a measure of semantic similarity it is common to use various kinds of vector similarity approaches. Conceptually this would typically be based on conditional distributions such as these:

$$P1 = P\ (\text{spacewalking} \mid \text{cosmonaut}) = 0.5$$
$$Q2 = Q(\text{spacewalking} \mid \text{astronaut}) = 0.5.$$

In an unary vector these properties would yield a perfect similarity. However, to be a valid vector analysis multiple properties would have to be taken into account.

1.9 Discussion

1.9.1 General Issues

In the discourse above we have briefly tried to describe various approaches to natural language understanding and its implications on natural language processing by means of computers. In here we have also tried to highlight the historic rivalry underlying the various approaches. The differences in view on linguistics are closely related to fundamental differences between schools of psychology. This controversy is closely associated with how psychologists view people and people’s behaviour. We may draw a line between the rationalists who look upon Man as an information processing entity where mind and memory plays an important role and where representation of the world is a prerequisite for successful capture of experience and language capabilities. Information is treated as discrete symbols and mental capabilities such as language processing, is thought to be largely inherited. The behaviourists will typically reduce emphasis on this and many will even denounce it entirely. Their belief is that people respond directly to stimuli exposed to them and that the proper response can be learned through repetitive exposure. Language capabilities are no exceptions. Many hold the view that first language learning is a consequence of exposure to speech in the child’s surroundings.

Our aim has not been to take stance with any part, but merely observe that these differences do exist. Moreover, we have carefully noted that there exists no general solution yet to the language understanding problem despite very optimistic promises and rash criticisms of alternative approaches through the last 50 years. In figure 6 we have included a list taken from Church (Church 91 with a few additions) that pins down the important divisions. It can be taken as an illustration of the discord. It is also important to note that some of the disagreements surrounding natural language processing are due to difference in goals and applications. The implications of pursuing language performance rather than language competence cause part of the distinction. People working with computers and computer understanding tend to forget that linguistics should also address the construction of sentences for eloquence in speech and to pin down grammars that enable written native languages as well as more efficient language learning.

Our approach is more that of an engineer and we propose the use of any technique that makes concept learning and information retrieval accurate and efficient for the purpose at hand. Even in the scientific community calls for unification of adverse theories and application-oriented work have been heard (Jacobs 1992, Ram 1999). Regardless of this it is important to understand that techniques applied for different parts of natural processing, from tokenisation through concept extraction may be biased to the views underlying the various “conflicting” approaches. This applies to various issues discussed here including acknowledgement of context upon understanding, the recognition of subjectiveness in analysis as well as the issue of discretisising information.

	Rationalism	**Empiricism**
Link to psychology	Cognitive psychology	Behaviourism
Well known advocates	Chomsky, Minsky	Shannon, Skinner, Firth, Harris
Model	Competence Model (Knowledge Model)	Noisy Channel Model
Contexts of interests	Phrase structure	N-grams
Goals	All and Only Explanatory Theoretical	Descriptive Applied
Linguistic Generalisations	Agreement and Wh-movement (see paragraph on X-Bar hypothesis)	Collocations and Word Associations
Parsing Strategies	Principle- Based, Chart (X-Bar), ATN's, Unification	Forward-backward, Inside-Outside[5]
Applications	Understanding "Who did what to whom"	Recognition Noisy Channel Applications

Figure 8 Summary of two basic approaches to natural language processing

Our study of the various types of literature discussed here indicate however certain preferences and tendencies. This is because one school has made a strong case or because there seems to be a growing consensus. In closing we will make a brief review of these.

1.9.2 Context

There seems to be a growing consensus that understanding and speech capabilities are strongly influenced by context and environment. We will not make any claim that language capabilities are not hard wired into people at birth. Yet it is clear that every spoken language is alive and develops in its own way within a community. Understanding the context is a prerequisite for understanding the use of words as well as their meaning. We have seen this in work that has been generally MOP based, ATN oriented as well as corpus based.

1.9.3 Expectation driven understanding

The work by Lebowitz that we highlighted in a separate paragraph used memories of historic episodes as a template for extracting information from text. The organisation of the memory and the concepts included are important aspects of the Dynamic Memory concept introduced by Schank. The conceptual representation is an instrument for language interpretation because it uses a part of the text to retrieve candidate memory structures to complete the interpretation process. In statistical based natural language processing we see a parallel to this through the use of n-gram models and conditional probabilities. The introduction of one word generates expectations of words that will follow. Expectations are generated based on structural relationships within the language and not between concepts within the world that the language addresses. Nevertheless it seems clear that experience and habits play important roles. This brings to bear ideas that understanding speech or text involves a complex feedback loop between new input and own

[5] Relates to direction of parsing

experience.

1.9.4 Subjectivity

Language understanding is subjective. This follows from the two other items discussed above. Situation and place determine context. People's experiences are never exactly the same. Since both seems to influence language understanding it is clear that subjectivity plays an important role. In addition we could have added the difference in goal and values. A person's objectives and both ethical and esthetical values govern people's perception. We believe that this will impose great constraints on any ontological oriented effort. Taxonomies of terms must be related to both context and the way the terms are used as well as to what they address.

1.9.5 Syntax versus Semantics

Over the past 30 years there has been an ongoing discussion on what focus to sustain, syntax or semantics, in order to pursue natural language processing. This discussion has surfaced in both the empiricist camp as well as the rationalist camp. In the former the answer is very much given. Yet we see a tendency that both are important. There must be significant interplay between syntax and semantic modelling. In fact the whole issue of language understanding circulates in our view around the process of communication. In terms of written expressions this implies that it is not sufficient to focus on the text alone, but on the whole system of author, author's knowledge, author's goals and constraints, author's world, chosen vocabulary, listener's context and listener's knowledge. The work of Ram et al. (1999 b) is a recent and quite impressive contribution in this direction.

1.9.6 Discretisation

In cognitive psychology symbol theory is very strong. There seem to be enough evidence to maintain focus on symbols. One reason is the language itself. Wolff points out that young children seem to discretesise language according to entities they observe. However work on collocations as well as semantic similarity shows that references to entities may correspond with a single term. In fact concepts may be completely without label. It may require several lines of text to pinpoint the nature of a concept. Moreover, a discrete term although well established may be rather fuzzy in its reference. One example is the bow and stern of a ship. On a slender sailing yacht it can be pretty difficult to determine the exact boundaries between the bow and the rest of the hull. Another example is the colour red. Where goes the boundary between red and pink? We may not take for granted that there is a crisp one-to-one connection between the real world entity and the referencing term although the concept and the reference may be quite unambiguous. Consequently we may have to look for rather long blocks of text or extensive patterns in order to address successfully a concept. Similarly, but worse is the problem of isolating the world in objects in order to define proper ontologies. We should also observe that tacit knowledge may not be a problem of incomplete dialectics, but an issue related to comprehension. It may seem that action oriented experience can only be really expressed through demonstration. In fact, it may seem at times that even the masters themselves do not consciously conceive the complex nature of a skill in depth. We should not take the ontological assumption for granted, but challenge it in order to get beyond the stage of simple taxonomies and aggregations. We pose here a bit of caution with the words of Dreyfus (1977):

"The ontological assumption that everything essential to intelligent behaviour must in principle be understandable in terms of a set of determinate independent elements allows AI researchers to overlook this problem. We shall soon see that this assumption lies at the basis of all thinking in AI, and that it can seem so self-evident that it is never made explicit or questioned." (pp. 206 – 207)

Our position is thus to keep a sustained point of convergence on symbols and tokens, but not limit ourselves to that. Work in both neural networks, pattern recognition, statistics (as pointed out for vector based similarity metrics above) as well as cybernetics have all shown that continuous variable approaches can do well and in fact, outperform symbolic systems for certain applications.

1.9.7 Challenges

In 1992 Paul Jacobs introduced his book on information extraction and retrieval (Jacobs 1992b). He identified four challenges. Although advances have been made we believe that the same goals are still valid.

- **Robustness**: Tomorrow's systems must do better what is already done well and not so well. Increased speed and accuracy is imperative. This must be combined with less domain-dependent knowledge. Techniques that are already robust such as parsing and morphology must excel further, while more knowledge-intensive techniques such as semantic analysis must receive close attention.

- **Retrieval strategy:** Retrieval must focus on information and less on documents. Text-based systems must address the broader issue of satisfying the information needs of many different systems and users. We should prefer direct answers to a query rather than a long text containing relevant information.

- **Presentation:** According to Jacobs there is a big problem with on-line text retrieval. People do not like to read. On-line text is even harder to read than printed material. Several new means should be pursued including summaries that combine different portions of text. Compressing text so that only key portions appear.

- **Applications:** A broader set of applications should be pursued. A broad effort will yield new landmark applications, but also feed back experience that can benefit theoretical work.

1.10 Information Extraction put into context.

Since the beginning of the computer era, research is conducted as to how computers can be used for tasks otherwise performed by human beings. The field of linguistics and its derivatives forms no exception on this curiosity. Already since the early 50's there is a shown interest in using automata for tasks otherwise reserved for humans. One of the earliest approaches where computers are used for an automated task including language processing on a larger scale is a project started in the mid 50's by IBM and aimed at automatic translation of Russian texts into proper English. Although rather unsuccessful, this enterprise thought the research community a few things. The first lesson learned is that there is more to translation as just cross-linking two dictionaries. Second, there was not enough knowledge about general grammars to be able to perform translation tasks while taking into consideration grammatical sentence structures. Third and last of all, there was no insight in the fact that even thorough grammatical knowledge would not solve all of the problems that are caused by the complexity of the problem. Only when more and cheaper computer power became available one started to realise that there was more needed for thorough language processing.

At the same time as the understanding of automated language processing started to grow, it was understood that many more applications could gain from linguistics. When the 60's showed an increase in implementations of statistical techniques and algorithms, and the "automata theory" plus Chomsky's theories on language became available, there was a starting interest to use computers for language processing in areas as diverse as story understanding, generation of world knowledge from existing text sources, automated generation of ontologies, hierarchy, concept structures, etc. Soon there where programs to be found that really dealt with language processing and could exist in spell-checking, hyphenation of words, using thesauri for word disambiguation and so on.

In many ways this can be regarded as the start of an era in which computers where regarded as automata being able of capturing world knowledge and reason with that. Initiatives as the CYC program (Lenat, 1995) aimed at capturing all available world knowledge), and emerging research fields like Computer Linguistics, Computer Vision, Machine Reasoning and Machine Learning all showed the growing confidence in the possibility to augment a computer with cognitive abilities. Computers where on the one hand used to evaluate theories (often in the form of rule sets) about language and grammars, on the other hand it was tried to go beyond that and come to a deeper understanding of the semantics of language. Based on ever more efficient sentence parsing programs, and using one of a wide variety of formalisms to describe grammars (Transformational Grammar, Augmented Transition Networks etc.), it was tried to analyse correct use of language often using some graph approach.

Nowadays such sentence parsers form the preliminary for nearly all language, information or knowledge based automatic approaches that perform one or the other task including natural language. Such parsers can often by classified according to the several levels of abstraction of a so-called Chomsky hierarchy. Since computers and information stored therein are often represented by texts, it is no wonder that language processing focussed on processing these corpora of potential information first instead of alternatives like speech recognition, understanding and the like.

There is a clear need for clarification of how terms as “information extraction”, “information retrieval, “ Natural Language Processing” are defined in the report at hand, because a variety of different definitions can be found in the literature. Figure 9 shows how the several processes relate.

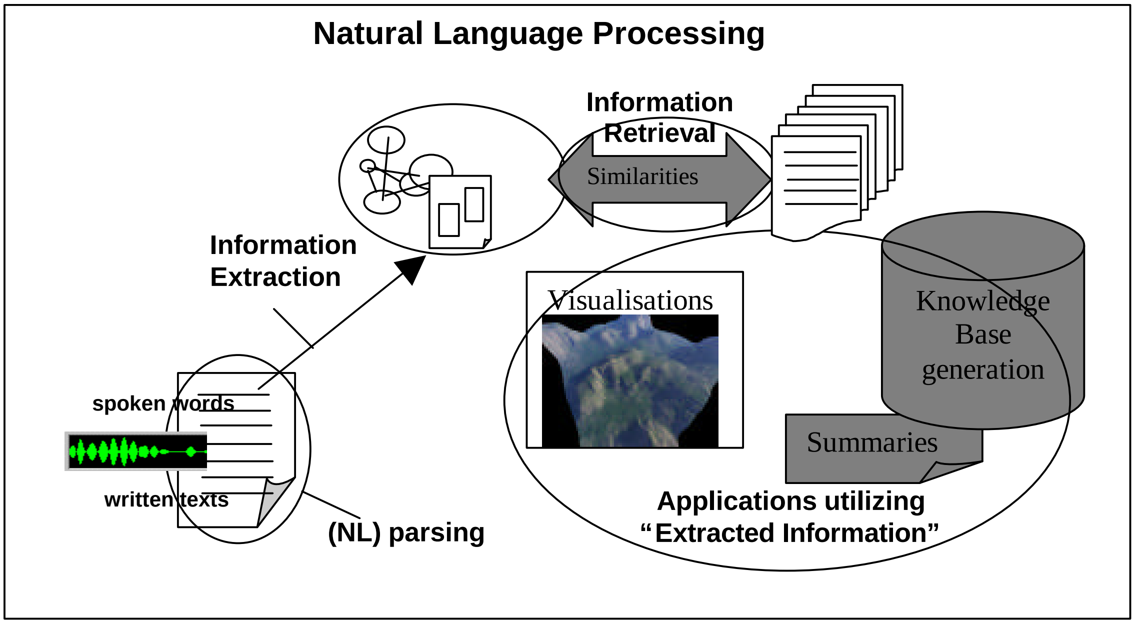


Figure 9: Natural Language Processing, NL parsing, Information Extraction and Information Retrieval

Natural Language Processing (NLP): is the analysis of natural language, be it the analysis of spoken language or written language, plus the application thereof.

Natural Language Parsing (NL-parsing): refers to the initial stages of the analysis process, where character streams are processed. The parsing process can contain all tasks that are possibly performed in the *lexical level*, the *morphological level* and the *syntactic level*.

Information Extraction (IE): is the process of extracting information from texts. Information Extraction typically leads to the *introduction of a semantic interpretation of meaning* based on the narrative under consideration. However, since Information Extraction also includes the "NL-parsing" stages, both IE and NL-parsing can point to the same processes.

Information Retrieval (IR): is the usage of IE results for retrieving information or documents from other sources. IR requires some *measurement or heuristics to estimate similarity* between the extracted information and other natural language (text) sources.

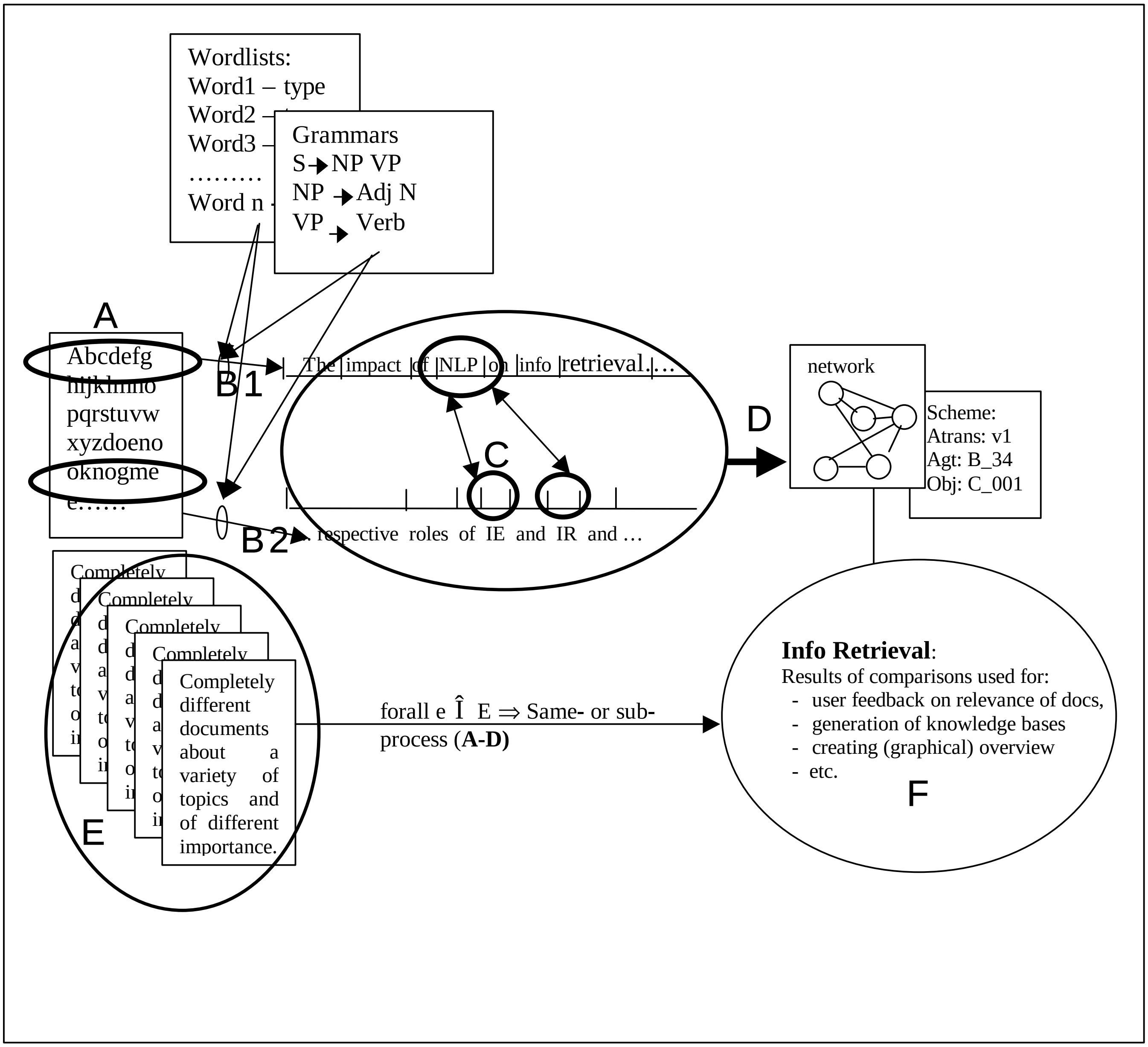

Wordlists:
Word1 – type
Word2 –
Word3 –
………
Word n
Grammars
S → NP VP
NP → Adj N
VP → Verb
A
Abcdefg
hijklmno
pqrstuvw
xyzdoeno
oknogme
e……
B1
B2
The impact of NLP on info retrieval. …
C
… respective roles of IE and IR and …
D
network
Scheme:
Atrans: v1
Agt: B_34
Obj: C_001
Completely
Completely
Completely
Completely
Completely different documents about a variety of topics and of different importance.
E
forall e Î E ⇒ Same- or sub-process (A-D)
Info Retrieval:
Results of comparisons used for:
- user feedback on relevance of docs,
- generation of knowledge bases
- creating (graphical) overview
- etc.
F

Figure 10: Information Retrieval in more detail

There is not much fantasy needed to imagine that the semantic interpretation that is generated from a particular document can be used in a larger variety of applications as IR alone. This is visualised in Figure 9 by the "application" tag. On a slightly more detailed level, Figure 10 describes which tasks typically are performed when starting the analysis of a document, text or information piece (A) that is analysed on a sentence level (B1, B2) using NL-parsers to identify sentences, paragraphs, words, special characters etc. and wordlists, grammars or rule bases in order to augment single words with information about the word type, place in the text, roles and possibly connections to other words in different parts of the text (C). The latter being tasks like co reference resolution, pronoun resolution (at the syntactic level), and the like.

- From the knowledge thus generated in phases **A** to **C,** some representational formalism will be used for representing the extracted knowledge. Often used are scheme-based representations or some kind of (augmented) network representations (see section on "representation of semantics"). Now, for performing Information/document retrieval, the representation that is thus generated (**D**) has to be compared against a set of other information sources/documents (**E**). Depending on the different approaches, this might be done using the original documents directly, or using some kind of document processing (i.e. subtasks equivalent to those performed discussed in **A-D**).

Typical applications of extracted information (cf. Figure 9 and Figure 10) could than be used for a variety of tasks, such as:

- *document* and *text* classification,
- user feedback on *relevance* of *documents*,
- *generation* of *knowledge bases* with extracted information patterns and semantically annotated information,
- *creating (graphical) overviews* of larger document sets from a specific, subjective viewpoint (f.e. a users' interest).

During the last 15 years, a significant number of approaches towards IE and IR (academic as well as commercial) became feasible (given the continuing increase in computing power that is available), necessary (given the connectivity of the 'online' community and the availability of 'electronic' information on the internet and within companies) and practical (opportunities to share and distribute information across people and communities increase at a pace unseen ever before).

In general, the several approaches towards language processing can be divided (as in many other sciences) into "fundamental" and "applied" linguistic research. According to (Neijt and Bakker, 1990), subdivision can be made of both fields into different topics (from NLP to Information Extraction). On the fundamental research side, one typically finds topics as:

- *lexicon programs*: that usually form the basis for parsers, since they contain information on single words and define their types, morphological appearances etc.
- *parsers*: make use of lexicons for checking input sentences on validity and grammatical correctness. Several parser types are found with a varying practicality.
- *model building (of grammars)*: building models of grammars is a necessary task if semantics of sentences are important or whenever the grammatical systems and their dynamics are reasoned about.
- *linguistic toolboxes*: research on grammatical principles is often supported by toolboxes that integrate the approaches mentioned earlier and make them available in one environment.
- *computer speech synthesis*: Besides generation of natural language texts, several disciplines in science deal with or are waiting for systems that can perform proper speech synthesis.
- *universal grammar and language acquisition research*: which is definitely one of the most

promising fields at the moment, with approaches towards automated learning of a new language given a universal grammar. There are a variety of problems that have to be solved, but generally there working approaches developed that even make use of machine learning techniques for inducing rules that hold within a certain language

- *grammar induction*: approaches are known that go beyond filling up and refining existing universal grammars. In such approaches it is tried to induce a complete grammar for a specific language, which can then be used for several automated tasks involving language.

On the other side there is the area of applied language technology and, as one might expect, many pragmatic issues come to bear here:

- *Optical Character Recognition*: prior to the analysis of text, it is sometimes required to read from printed pages or hand written text. Financial institutions and post offices are the major parties concerned if it comes to large non-electronic text reading tasks. Together with producers of copy machines and scanners, they are the major driving force behind this type of technology.
- *keyword based technology* refers to a variety of techniques ranging from the well known spell checkers, hyphenation programs, word based search and translation services. Another service such keyword technology might deliver is encyclopaedia look up.
- *rapport generators*: a quite natural application area for natural language synthesis programs is the generation of reports. Currently there are completely automated applications reported that can analyse data that stems from a supermarkets' cassier systems, perform automatic data mining on that data and write out a monthly report in natural language while including the latest findings. Other companies use natural language synthesis programs for the generation of understandable reports from formal models that will be used in expert systems (Kristen, 1993).
- *natural language DB interfaces:* several academic and commercial research institutes have conducted and are conducting research into the topic of defining natural language interfaces to databases. Possibly the most well known for the moment are search engines on the internet (f.e. AskJeeves) that claim to have a robust natural language understanding interface to their search engine data bases.
- *automatic content analysis*: can be scaled on a range from shallow to deep content understanding. Shallow are those approaches that only take into consideration keywords and possibly combinations of those, whereas the deeper the "understanding" of a text is, the more one will need to build upon a complex structure representing the actual meaning of a text. A good example of such a richer and rather complex structure building formalism makes use of graphs, and is proposed by (Schank, 1977).
- *automatic translation*: is a topic that formed the basic for modern computer supported language research in many respects (Remember that one of the earliest projects with NL at IBM was in this direction). For profound translations, a complete NLP process is required including stages where there is a syntactic, morphologic and semantic analysis stage. This is needed due to the fact that translations are often very sensitive to the choice of words in relation to a certain context (defined by other words and interrelations). The importance of a proper semantic representation becomes clear when translating sentences containing sayings. As an example, the dutch saying "Wie A zegt moet ook B zeggen" literally translates with: "Whoever says A should also say B", whereas a proper translation on semantics will have to translate with: "In for a penny, in for a pound".

1.11 Classical approach to Information extraction and text understanding

Where the basics for all language related processing should be sought in computer linguistics, information extraction is the logical next step for utilising the collected knowledge about words, word types, cases, senses and similar information. From the research topics discussed above, it can now be seen that the last three mentioned (i.e. "NL interfaces to DBs", "automatic content analysis" and "automatic translation") are dealing with something that can be regarded as information extraction. In all three of these tasks a semantic structure representing "meaning" or "context" is required for a good performance on the respective tasks.

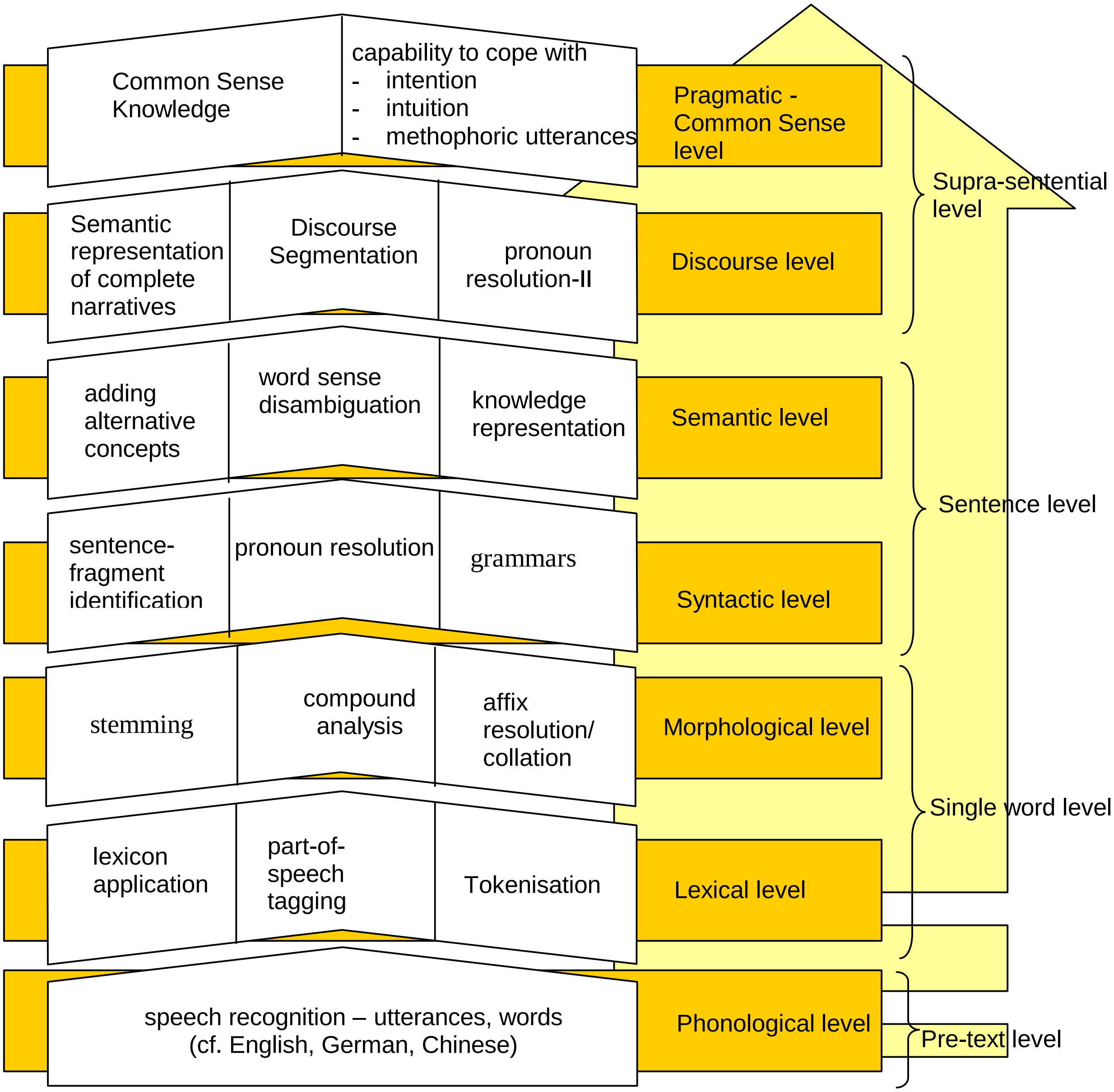


Figure 11: "Classical" Natural Language processing decomposed.

All the other tasks that where mentioned basically deal with either the generation of machine readable texts or with the preliminaries to be able to analyse sentence structures and make purely syntactical and morphological analyses of texts[6]. At this point it should be said that the difference between the natural language parsing phase on the one hand and the information extraction phase

[6] Of course such "contextual" knowledge might be helpful in some of the other cases as well, but we do not regard this as absolutely necessary. Reports, for example, can be generated using a simple grammar or even more simple using schemes/templates, where the lack of such contextual knowledge is not necessarily a problem.

on the other hand really is vague. It might be helpful to regard all operations on the original text string as such (single word recognition, finding word types, performing a morphological analysis, generating a sentence representation based on a grammar) as belonging to the parsing component. This automatically means that all operations that are dealing with the further application of this parser generated knowledge belong to the information extraction phase.

The generation of semantic knowledge in the information extraction phase bases on the results of the parsing steps that can be of varying “analysis-depth”. Some approaches build on knowledge about word types only, whereas other approaches go beyond that and require a deep understanding of the sentence structure before being able to generate the semantic representations that are aimed at.

Currently, when analysing a variety of published IE approaches, the whole range can be found, from utilising simple word list (often augmented with frequencies that are learned on documents corpora), through hierarchies of concepts that are related to each other by specific relation types (Concept-SuperConcept relations, Part-Whole relations, etc.), towards even richer semantic representations like the theory of Conceptual Dependencies (Schank, 1977). Generally speaking, such representational schemes are believed to represent a world state, beliefs or factual knowledge that is implicitly present in the documents that are analysed. There are very diverse opinions on how semantics are represented. Some research groups see semantics of meaning as the semantic as defined by the semantics of some logic representation. Others propose that context is described by statistical properties over text characteristics (like (normalised) word frequencies, the appearance of a variety of words as a function of a texts’ length, etc.). Finally, there is a group of approaches in which semantics are described using AI techniques. Popular representation schemes are semantic networks and connectionist models, but also ontological and hierarchical representations can be found.

In *classical natural language processing* approaches, a typical build up through a variety of levels for NL processing exists. However, despite this common understanding of the analysis process, differences in the finally used technologies exist between (language-specific) approaches.
Textbook discussions on Natural Language processing typically report 7 levels of abstraction (listed from high abstraction levels to low abstraction levels):

- Pragmatic or Common Sense level, the level of “real world” knowledge.
- Discourse level,
- Semantic level,
- Syntactic level,
- Morphological level,
- Lexical level,
- Phonological level
-

Figure 11 provides an overview plus context for these seven levels, whereas every single level is decomposed in its main phases. First of all, the figure represents an ascending level of abstraction from its bottom to top. Where the processes and techniques at the bottom of the figure are relatively well understood and well researched, the processes that take place at the top (if at all) are often based on weaker theories. For this reason, and due to a lack in common terminology when approaches are described in general, it is sometimes only possible to derive conclusions based on circumstantial evidence. Therefore it is not always clear in how far a certain tool performs a specific action, naturally all kinds of gradations can be found (i.e. the extend in which proper nouns are recognised in a text). In the following, a discussion is provided on these 7 levels of abstraction from the lowest level of abstraction (phonological) to the highest level of abstraction (Pragmatic, Common Sense).

1.11.1 Phonological level

The phonological level is not found in so many NLP systems. It is at this level where speech utterances are transformed or interpreted such that they can form the input to the morphological level. Often at this level sounds are transformed into some textual natural language representation.
A variety of techniques that can be used at this phonological level are available commercially or as free/shareware[7]. Most common is the use of connectionist models to come from differing, fuzzy and chaotic speech patterns (differing across and among individuals) to a standard uniform interpretation in terms of written language. Such systems typically come with a pre-trained neural network that has to be customised by the end-user of the system. Only in this manner the systems adapt to an individuals accent, pronunciation and voice peculiarities.
Since a trade-off between the complexity, quality and runtime of these systems has to be made, combined with the fact that it has shown to be a rather difficult task to recognise words from utterances alone, these systems are normally augmented with a dictionary and/or a thesaurus. Combing connectionist/probabilistic output with the most likely word from a dictionary or thesaurus gives such systems a performance that is good enough for them to be employed in real life practice. Some systems even make use of algorithms on the syntactical level, and try to predict what the most likely following word will be given a certain sentence structure and the captured utterances so far in order to even further enhance the quality of the speech recognition process.

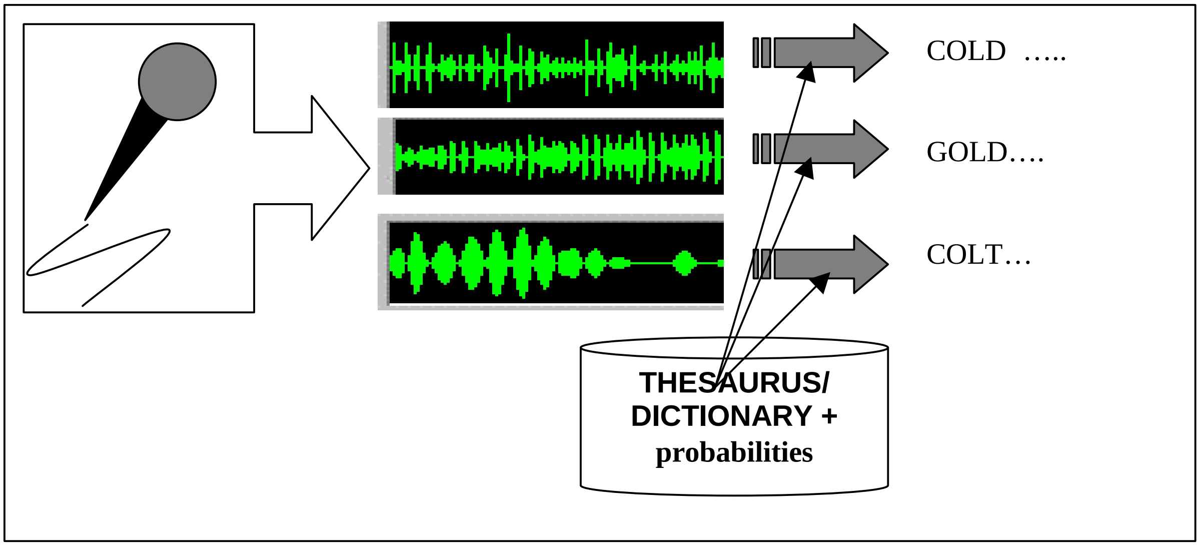


Figure 12: Translation of speech into written language.

Well known in this respect is the usage of "Hidden Markov Models", a technique with a firm statistical basis. HM models typically consist of two phases, the first transforming the sound waves into a graphical (orthographic) representation (cf. Figure 12), and the second using a dictionary/thesaurus in combination with (corpus based) information about joint distributions of words within texts.

1.11.2 Lexical level

As a textual representation of a communication is available, a *tokenisation* phase is performed in which single words, delimiters, full stops, special characters etc. are identified in a text. Such an identification is naturally easier from (more formal) written text as it is to identify sentence structures from spoken text. For reasons of simplicity we will restrict ourselves to those scenarios where a written document or text is available for analysis. Tokenisation forms an important basis

[7] Examples of such speech recognition and translation software are: Dragon NaturallySpeaking (www.naturallyspeaking.com, IBM's ViaVoice (Freely available for Linux at: www.ibm.com) or Philips' Freespeech (www.philips.com).

for the processes taking place at the higher levels of abstraction, such as (proper) noun identification (and possibly sub-types thereof), compound analysis and sentence fragment determination. During tokenisation, typically a database is generated containing the token types, offset/position in the text and additional information about the tokens for later use. The role of the lexical level is to determine the word categories that words belong to, as well as identification of specific word types that do not necessarily follow from the application of a grammar (such as proper nouns). Generally two word groups can be identified on the lexical level, *grammatical* words and *lexical* words (or *function words* and *content words* respectively, cf. (Winograd, 1972)):

- *Grammatical words*:
    - Articles (the; a; and),
    - Adjectives (green; white; cold),
    - Predeterminers (half; both),
    - Prepositions (under-;over-;un-) and
    - Pronouns (it; its; he; she)

- *Lexical words*:
    - Adverbs (shortly),
    - Qualitative adjectives (larger;smaller),
    - Nouns (project; people),
    - Verbs (co-operate; communicate)
    -

The word groups that bear the real semantics of a text are identified as being the *adverbs*, *adjectives*, *nouns* and *verbs* (cf. (Winograd, 1972) (Allen, 1994)). As a consequence you will find that many approaches that deal with the analysis of documents and corpora of texts will concentrate on one or more of these four groups. Depending on how elaborate an approach is with respect to the understanding, it will typically include the analysis of nouns, adjectives, verbs and adverbs respectively (in this order).

The grammatical words represent a group of words that is also referred to as “closed class words” (since there are hardly new words added to this class) and they are easily captured in dictionaries for (table-) lookup. The second category of “open class words” poses a problem for table lookup or rule based identification. On the one hand the number of lexical words is not stable, i.e. new words belonging to this class are added to the language regularly. On the other hand, besides of the regular introduction of new words, there is the problem of the several tenses of verbs and the modularity of words (plural, singular), which often makes them hard to identify.

Often approaches solve these problems by using direct table lookups for the grammatical words, and using a more or less knowledge intensive approach for the lexical words. For the latter, rule bases can be used (with as main disadvantage their maintainability), or statistical heuristics can be applied for calculation of the most probable word types given a certain sentence fragment and a corpora of texts (cf. Hidden Markov Models). Often used in this respect ate POS-Taggers (Part Of Speech Taggers). Results of such taggers are frequently published at MUC conferences. Statistical POST calculates conditional probabilities (using Bayes theorem) and uses actual appearances of word sequences to calculate the most probable word type for a particular word. Whereas in theory such a probability may depends on all preceding words in a text, in practice POST (and other systems based on probabilistic theory) take a pragmatic approach and calculate the conditional probabilities according to the *n* preceding words. This is then referred to as “n-Gram Models”.

On top of this, often a separate heuristics is implemented to identify proper nouns. Once identified proper nouns can be sub-divided into categories like person, organisation, location, etc. (cf. the SPPC system (Declerk, 1998)). Such rules build upon the results of the tokenisation process. Basic rules for identification of proper nouns are often defined in the following way:

- if *all_capitalised* then proper noun,
- if *not the first word in a sentence*, and *first letter is a capital* then proper noun (check lists with names and locations for possible proper noun types.),
- if *string contains "& Co." or "ltd."* then tag previous word as proper noun (type: company).

Words that are thus annotated according to the processes described above are used in typical information extraction tasks. Often a special focus is on proper nouns in combination with the categories of words mentioned above. The knowledge extracted here can now be forwarded to the next level of analysis.

1.11.3 Morphological level

After the lexical analysis, it is necessary to conduct *co-reference* resolution. Co-reference resolution is important in that it aims at finding as many related appearances of a certain concept in a text, a problem that is sometimes difficult to solve due to the large number of different ways to refer to a concept. The reason for this is that many information extraction systems are based on probabilistic approaches, and therefore need a proper representation of such co-occurrences of concepts within one text, often using slightly different tokens. In total, three types of *needs* for co-reference resolution can be identified:

1. Co-references of several tenses of *nouns* and different modularity of *verbs*, cf. The words "carrier", "carriers", "carried" and "carries".
2. Co-references based on *similar concepts*, using different words, (cf. the use of the words "car", "automobile" and "vehicle" to refer to the same entity within a certain context),
3. Co-references based on similar *named entities* that are referred to using several proper nouns and abbreviations within a document (set).

Whereas the second and the third co-reference resolution needs are covered on the lexical and semantic level, the first need can be solved at the *morphological* level. Often this problem is tackled through stemming and affix resolution. During stemming, the several tenses of a verb are reduced to their root, in order to be able to compare relationships in a given (con-) text:

WordEnding(verb) is {-ing, -s, -ed} **then the verb root is** *word – WordEnding*.
(e.g. word:grasps -> root:grasp)

Plural forms of nouns are often reduced to their singular forms for the same reasons:

WordEnding(noun) is { -s} **then the word root is** *word – WordEnding.*
(e.g. word:cars -> root:car)

At the same time, it is often necessary to define rules or heuristics for finding the "basic" word forms of *derived* words (*derivational morphology*). Typically affixes (prefixes and suffixes) have to be resolved. Basic approaches can make use of simple heuristics for performing such tasks.

	Pre stemming word ending	After stemming word ending
1	*ly	*
2	*ity	*
3	un*	not(*)

Table 2: stemming and affix resolution

The basics of stemming are relatively simple and can often be performed by using a simple rule base. Typically it encompasses the transformation of nouns, verbs and adjectives into their basic stem and tenses.

The other task that is performed on the morphological level is the decomposition of compound words. Compound words are not so common in English (although words as "database", "hardware" and "motorcycle" theoretically would fall into this category), but play a significant role in germanic languages like German, Dutch and Norwegian (to name a few).

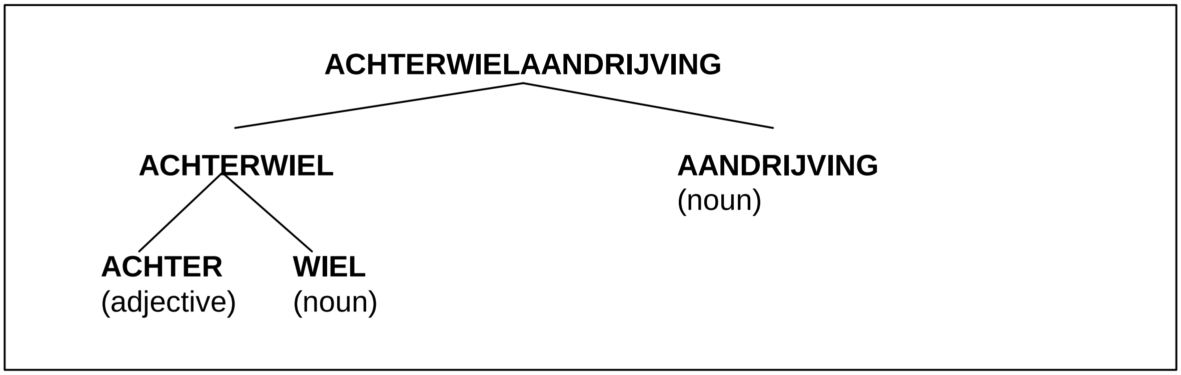


Figure 13: decomposition of a compound noun.

Co-occurring words normally refer to a specific, detailed concept, and are therefore a focus for linguists aiming at capturing the essence of a text. Whereas in languages as English the challenge is to identify such co-occurring words, in language like German the opposite is true, and the challenge in this case is to recognise composites and identify the words such a composite is made of. The identification of co-occurring words based on evidence or distance metrics requires analysis on a supra-sentential level, and is therefore not part of the *morphological level*.

The *decomposition* of compound words, on the other hand, is a topic that is taken care of on the morphological level. It is a process that takes place at the same level as the earlier mentioned *stemming* process. As an example the word "achterwielaandrijving" (taken from Dutch: "rear wheel drive") can be decomposed in three consecutive words: *achter* + *wiel* + *aandrijving* (cf. Figure 13). Here the main noun refers to the *drive* (or the fact that something is *driven*), whereas the second noun determines the object that is driven, namely a *wheel*. The adjective finally tells us that it is not just a wheel, but specifically the *rear wheel* that is driven. It is easily seen that a thorough analysis and appropriate representation of compounds will lead to a better understanding of the whole text. Usually compound analysis provides the means for a richer context description by delivering additional information about single nouns.

However, such simple approaches are often not sufficient for a perfect analysis of derivational or inflectional morphological analyses since the context of the word (grammatical (sentence level) semantic and discourse level, cf. Figure 11) is also important in the analysis. Since there is no solution to be found at the sentence level, it is often the case that multiple analysis results are propagated upwards in the process and decide upon the correct analysis on a higher level of abstraction.
At the next higher level (the syntactic level) the analysis of the words that make up a text is elevated up one abstraction level. At the syntactic level words are put into relation to each other and thereby dissolving some of the unclarities that still exist in the interpretation.

1.11.4 Syntactic level

At the syntactical level there are basically two important tasks to be performed. On the one hand, there is the task of identifying the constituents (or sentence fragments) that a particular sentence is composed of on the other there is the task of assigning the exact roles to individual words, taking into consideration the grammatical rules of a language and a description (grammar) of how words can be put together in that language. In order to be able to compute such a constituency analysis, a *grammar* is needed, as a formalised theory on sentences that are allowed in a given language, and a *parser*, which is used for the analysis of a particular sentence against such a grammar.
A variety of grammatical formalisms exist, where the more well known grammars are *Context Free Grammars* (CFG), *Tree Adjunction Grammar* (TAG), *Generalised Phrase Structure Grammar* (GPSG), *Dependency Grammars* (DG), *Systemic Grammar* (Halliday, 1970) (Winograd, 1972), *Case Grammars* (Fillmore, 1968), *Recursive Transition Networks* (RTN). CFG, TAG and GPSG can be regarded as representatives of unification grammars (or constituency grammars), aiming at a recursive decomposition of sentences in meaningful fragments (cf. Figure 14). In these grammars the syntax of a grammar is described according to predefined categories (S:sentence, VP:Verb Phrase, etc.), which stand in a part-whole relation to each other. Such CFG can often be described using a BNF notation. Context Free Rewrite rules and Recursive Transition Networks form other typical grammatical formalisms that are especially suited for capturing recursive description of sentences and the parsing thereof.

According to (Zarri, 1997), several problems are related to the classical analysis of language using Unification Grammars. Whereas grammars can be used for relatively deep analyses of natural language based on formally correct language, problems arise when applying the grammars to less formal texts, like those retracted from spoken language (where many utterances are rather "ungrammatical" although they have a clear semantics) or text written by non-native speakers. Furthermore there are certain specific problems with the application of such unification grammars to texts. For example, dealing with "degree expression" (Staab, 1999) or "distributive" structures[8] (Zarri, 1997) are not possibly generated using level 2 or 3 grammars[9], whereas level 1 grammars can using a representation with a complexity that puts into question its usability in real world applications.

[8] Distributive structures are introduced by adverb like "respectively" and are problematic for representations using level 2 and 3 grammars, as stated in the text. (e.g. "Oslo and Bergen are the capital and main cultural city of Norway, respectively").

[9] cf. Chomsky's four level classification model for unification grammars. Levels range from 0 (no constraints on the form of the head and body in rules) to level 3 (severe restrictions on productions). Level 2 and 3 represent the "context-free grammars", level 1 the "context-sensitive grammars".

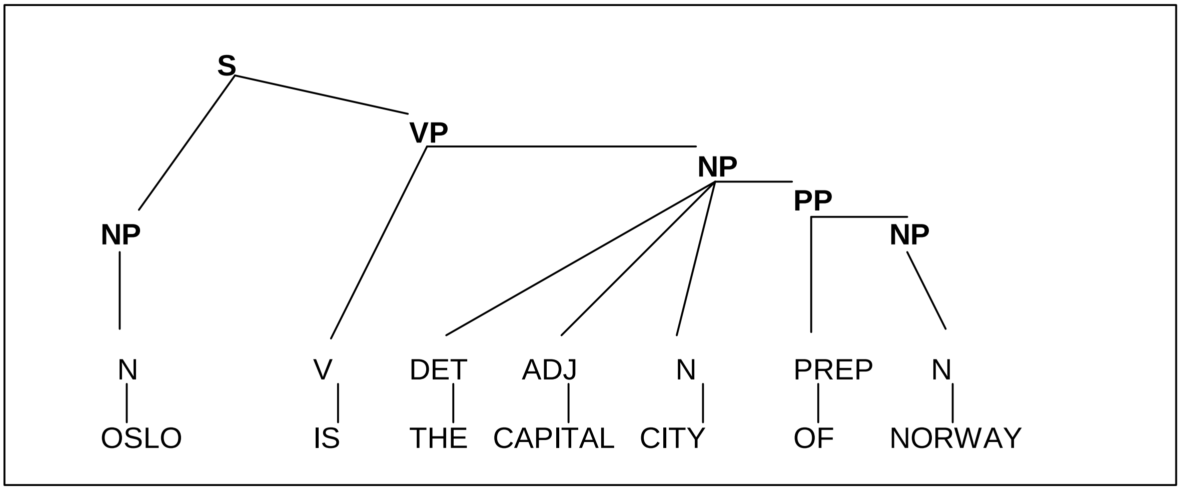


Figure 14: A typical example of the application of a Phrase Structure (constituency) grammar.

Dependency Grammars are used as a much more comprehensive notational method for grammatic knowledge. A good description of Dependency Grammars can be found in (Hudson, 1990) (Bröker, 1999) and an application of these Dependency Grammar can be found in the ParseTalk system (Hahn and Strube, 1996) (Staab, 1999). Typical features of DG are the fact that words are binary related to other words (cf. Figure 15).

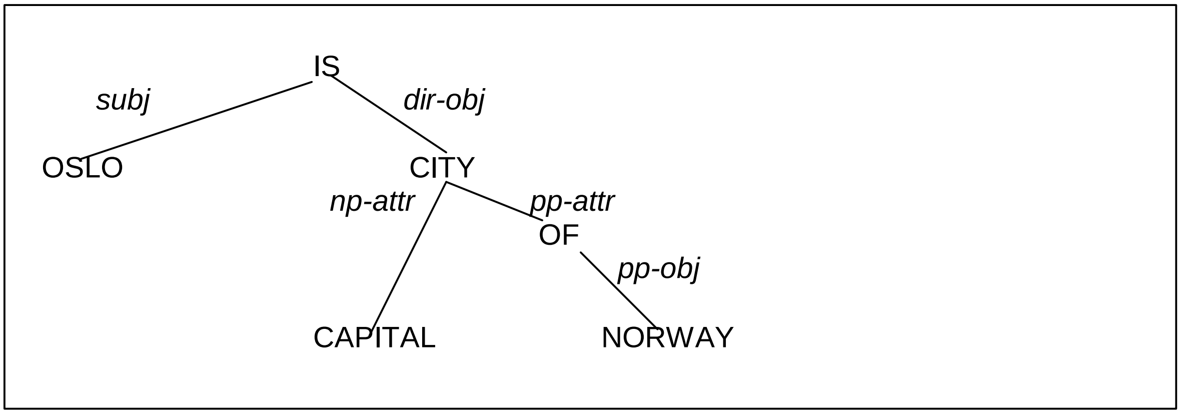


Figure 15: Describing a sentence with a Dependency Grammar (DG).

Relation types or restrictions that words can have to other words are *valencies* and *vacancies*. *Valencies* "describe the standard argument structure of a head word and allow for establishin relations between the head word and (specific) modifiers… *Vacancies* of a word describe requirements that have to be fulfilled by the head word in order that a dependency link may be constructed….(Staab, 1999)". Establishing dependency relations is allowed as both restrictions that stem from valancies and vacancies are fulfilled. *Dependency Grammars* have several reported advantages that make them preferable for information extraction tasks above the use of *unification* grammars ((Staab, 1999):

Syntactic Simplicity: only those syntactic entities that appear in the texts are used, no additional entities are postulated,

Semantic Correspondence: Syntactic and semantic links closely correspond,

Lexicalisation: DG approaches are fully lexicalised while valency, vacency and word order descriptions are attributed to words,

Discontinuity: Long distance dependencies can be captured more easily.

Text Grammar: The notion of "dependency relation" easily carries over to the level of coreferential words and anaphoric relations between those.

For reasons mentioned above, Dependency Grammars can be seen as a means to find and represent a system consisting of *concepts* and *roles*, based on some natural language textual representation. Combined with the fact that DG can be represented in Description Logic (cf. Staab, 1999) which has a model theoretic semantics and thus counteracts one of the former disadvantages of DG (namely its lack of formal treatment). Furthermore it is necessary to realise that DG approaches like the one referred to are using grammatical knowledge captured in an (obligatory) lexicon (Bröker, 1999). The knowledge contained in such a lexicon consist of a *world class hierarchy*, capturing restrictions on the syntactic roles of a word, and the *lexeme hierarchy* capturing features that are associated with word stems (e.g. the noun " carrier" and the verb "carries" have in common certain features that are attached to the specific word, cf. the *stemming phase* at the *morphological level* in figure 5).
Related to the dependency grammars are Case Grammars, which are often used in AI for description of "context" in computer internal "world" representations (Fillmore, 1968). Schank's Conceptual Dependencies (Schank, 1975) might also be regarded as a variant of such *case grammars*.

It is now possible to go from the very formal unification and dependency grammars to a relatively "new" approach to natural language parsing that is much more based on the capturing of *ad hoc* context descriptions and relations based on a different paradigm. In such approach (sometimes referred to as "Data-oriented Processing, DOP), parsing and representation of meaning should not be depending on a non-redundant set of formal rules. On the contrary, according to this doctrine, parsing should completely take place based on probabilistic processes, parsing new input (texts) by analogy. Doing so has as advantage that no lexicons have to generated for each and every domain that might be encountered, and in addition to that, that parsing is not longer depending on one particular view on what is the right set of non-redundant formal rules for parsing. In the DOP approach, a parser uses *corpora* of data that represent (historical data within) a particular domain. Meaning is now captured within structures that base on the information that is generated at the lexical and the syntactical level. In the eyes of some, such approaches are much closer to an answer to the remark made by (Winograd, 1972), where it says:

"What is needed is an approach which can deal meaningfully with the question "How is language organised to convey meaning?" rather than "How are syntactic structures organised when viewed in isolation?".

-- *Terry Winograd*

Although this paradigm is not too wide-spread for the moment, a variety of systems is available that make use of probabilistic principles for natural language parsing, and the idea of capturing semantics and meaning with different, experience or heuristic based techniques seems promising.

1.11.5 Semantic level

At the semantic level knowledge representation level, word disambiguation and or the introduction of new, alternative concepts in the representation of a text take place. It is important that a representation of semantics allows for reasoning with this knowledge, and provide a sufficient representation of the context in which a specific word or word group should be understood and can be "disambiguated". So now, based on the previously generated syntactical structure as well as the semantic meaning assigned to the individual words, it is necessary to put it all together in a "meaning of the sentence". Four topics are identified to play a role at the semantic level: *knowledge representation*, *word sense disambiguation* and the *extension of knowledge representations* with synonyms or related words.

The process of adding semantics to a representation of some natural language text requires a sequence of steps. It is important that a sufficient *knowledge representation* formalism is defined. Widespread in the literature evaluations and assessments of approaches are found that all come with advantages and disadvantages of their own. Such approaches rang from pure higher order logic approaches to very implicit knowledge representations (down to the level of modelling and representing worlds in DNA-strings or highly connected networks). It is not the goal of the current deliverable to provide an overview over the current state of the art in knowledge representations[10]. Instead a short overview is given of approaches for capturing semantics from syntactical knowledge. These approaches are listed in a sequence that reflects their increasing semantic content. Approaches to natural language text interpretation often are divided into *shallow* and *deep* semantic representations.

The shallowest of the semantic approaches is possible the approach in which semantic procedures are used for augmenting syntactic analysis. This refers to nothing more as a process in which the syntactic analysis of an input sentence leads to several representations of the input pattern. These representations just exist next to each other, for example as representation in "Quasi Logic Form"[11] where a set of attachment rules is used to decrease the number of possible readings as far as possible.

The approaches on the next higher abstraction layer for representation of semantics are the so-called "Case Grammars". Such grammars are based on the observation that verbs (as identifiers of "functions") can be classified in specific groups, all with their own requirements on the "slots" that are connected to a verb of that particular class. As example, consider the following structure where a case for a general transitive verb is generated:

<SUBJ> <TRANS_VERB> <OBJ>

In such cases, the analysed sentence that contains the transitive verb should also contain a noun or proper noun that has the subjective case. This noun can then be included in the (SUBJ) slot. Case grammars can contain several of such frames, f.e. for transitive verbs like in the example, but also for the intransitive verbs and can be different for active and passive sentences. The cases that are identified through the application of such frames can be used for knowledge representation, where trigrams can be generated based on this knowledge, or the output of case grammars can be used for the construction of conceptual graphs. As extension of such case grammars one can consider semantic case classifications (cf. Noble, 1988). In this kind of grammars, a variety of roles is assigned to cases, thereby providing a classification. Proposed classification categories include *agent*, *experiencer*, *instruments*, *results*, *goal* (as in location), *source* (as in location), *counter_agents*, *patient*, *force*, *etc*. Semantic knowledge is needed to make the categorisations

[10] Quite on the contrary, there will be a separate deliverable of EU project 10132: OntoKnowledge that is dedicated to the issue of representation formalisms.

[11] For an example of QLF and a mor elaborative introduction into it, refer to (Zarri, 1997).

based on the original role of a specific word in a sentence, hence the name semantic case classifications. As example, consider the following two sentences:

A) Jack hits Bob.
B) Tracy rebuilds a car.

In sentence A) “Jack” is the *agent*, “Bob” the *experiencer*, and they are connected by a *transitive verb*. In sentence B), “Tracy “ is the *agent*, while “car” is the *result*. It is easily seen that a semantic case grammar is hardly generally applicable, and a universal case grammar is not to be expected. Semantically augmented case grammars have to be so extensive that they can capture all the different semantic categories in relation to other knowledge in the sentence (f.e. knowledge about the function of a verb and the “cases” that are allowed for a particular verb). One of the biggest challenges for case grammar approaches seems to be the fact that words can have so many different meanings that a similar grammar frames is often incapable of making the right distinction between them.

As an answer on these shortcomings, a logical and necessary extension on the semantic case frame approach was defined. In these, generally known as *frame-based* approaches, focus was not on a particular input sentence, but instead on a complete discourse. Doing so has the advantage of just having to identify the main roles and actions in a text, thereby providing a *condense* representation of a larger text. Frames are typically specialised on their applicability within a certain domain, a well known example domain being that of terrorist attacks (Ram, ??). During the syntactic parsing of a text, all clues that refer to values for specific slots within such frames are used for filling these slots. In the course of parsing a text there is an incremental build up of semantic knowledge about a specific action within the discourse that is analysed. Perhaps the most important disadvantage of this type of semantics is the fact that they require the discourse at hand to be known in order to be able to fill the different frame slots with knowledge.

Whereas frame based approaches lack the knowledge and representation possibilities needed for capturing causality, order and spatial and timely distribution, this lack is (partly) overcome by the introduction of scripts and plans. These are used for the understanding of stereotypical situations, like going to the pictures, or visiting a restaurant. Such scripts describe roles (who is typically involved in the store), constraints on such roles (a cinema possesses one and only one unique location, a ticket vendor is either a person or a dispenser, etc.), preconditions for the scripts to be applied, effects of the application of the script (a content feeling after visiting a restaurant, an emotional mood after visiting a movie, etc.) and possibly a number of sub events that take place during the time span of the script. Such sub events can naturally be scripts of their own. In such scripts, the observed presence of the preconditions triggers the possibility for the application of a particular script. Higher order scripts can also be regarded as plans, and thereby increase the possibilities for a right interpretation of a particular text. If an appropriate plan is identified (i.e. about Peter who has as goal to cross water (the channel)), then the interpretation of a particular text part (Peter has a course on canoeing. The teacher tells them about weather forecasts and takes them out) becomes much easier. The overall plan might be to cross water, a plan that can be subdivided in choosing a particular type of boat, learning how to use it, learning about specific weather conditions, and finally the crossing itself. The goal of learning about specific weather conditions can be described by a script that subdivides learning in theory and practice, etc. Scripts and plans are pretty natural ways to describe stories and courses of action., but share the problem of how to cover all possible real word situations, and how to choose among a practically infinite number of such scripts and plans.

While a more general theory for representing meaning was needed for the representation of universal domains, more sophisticated approaches appeared. One rather influential theory was the theory of Conceptual Dependencies (Schank, 1977). In this theory, an explicit classification was made in *mental constructs*. Instead of dividing and classifying the world according to predefined plans or scripts, it is also possible to classify concepts, actions and states on an abstract level. This allows for building categories in which each and every action possible in the real world would fit, and thus provides the generality that is needed to deal with new, unknown domains. In the theory of Conceptual Dependencies it done exactly so, the world and communications about it can use a variety of constructs that can be divided in *acts*, *picture producers* and *states*. Together they are sufficient for the description of arbitrary worlds.

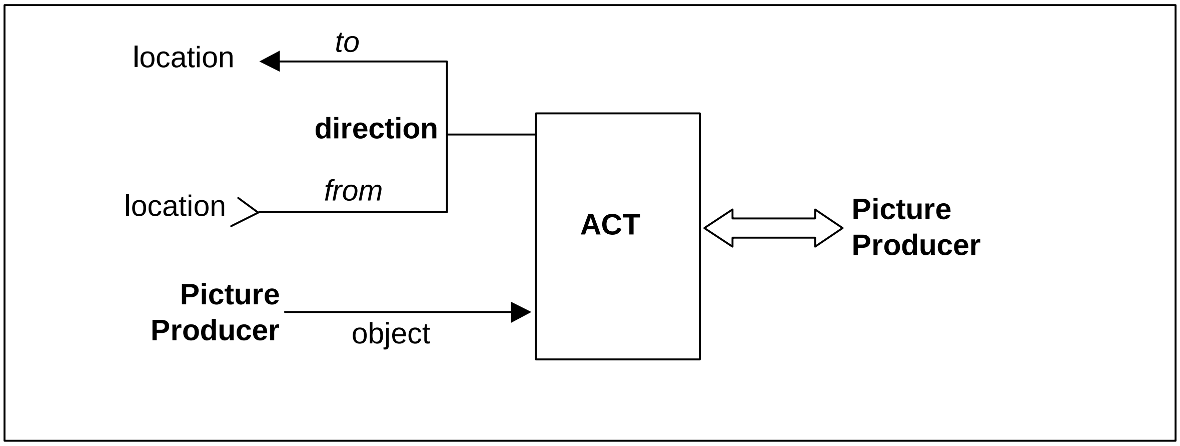


Figure 16: Describing a sentence with a Dependency Grammar (DG).

The theory of Conceptual Dependencies differentiates between a finite set of primitive acts (eleven primitive acts where identified in the initial publications, this set evolved slightly in later theories). Each act describes a transfer (or moving) of physical or mental objects. An actor acts upon such objects and the movement has a direction (from, to or none). Slots describe these acts. Verbs consist of one or more of such *primitive acts* that constitute the lowest level, i.e. they are non explainable. One or more of such primitive acts can, possibly using predefined slots, describe specific verbs. The twelve primitive acts that are consistently found in the theory of Conceptual Dependeny are[12]:

1. MOVE: refers to a physical movement of an object from one location to another.
   Slots: *agent*, *object*, *direction*(*from*), *direction*(*to*).
2. ATRANS: transfer from one “abstract” state to another (i.e. selling property rights).
   Slots: *agent*, *object*, *direction*(*from*), *direction*(*to*), *instrument(PTRANS, MTRANS, MOVE)*.
3. PTRANS: physical transfer of an object (i.e. granite rocks from a delving pit to a house).
   Slots: *agent*, *object*, *direction*(*from*), *direction*(*to*), *instrument(PROPEL, MOVE)*.
4. MTRANS: mental transfer of ideas between persons.
   Slots: *agent*, *object*, *direction*(*from*), *direction*(*to*), *instrument(SPEAK, ATTEND)*.
5. MBUILD: refers to the construction of mental ideas or plans.
   Slots: *agent*, *object*, *direction*(*from*), *direction*(*to*), *instrument (MTRANS)*.
6. PROPEL: a moving act, where the agent and the object are different.
   Slots: *agent*, *object*, *direction*(*from*), *direction*(*to*), *instrument(MOVE)*.
7. EXPEL: expelling material from the body. (Could be seen as a MOVE act)
   Slots: *agent*, *object*, *direction*(*from*), *direction*(*to*), *instrument(MOVE, PROPEL)*.
8. INGEST: intake of material into the body. (Note: this could also be seen as a MOVE act)
   Slots: *agent*, *object*, *direction*(*from*), *direction*(*to*), *instrument(PTRANS)*.
9. GRASP: to enclose an object in such a way that it can be manipulated by the agent.
   Slots: *agent*, *object*, *direction*(*from*), *direction*(*to*), *instrument(MOVE)*.

[12] See also paragraph 1.4.3.1 to 1.4.3.4 on Schank’s Conceptual Dependencies.

10. SPEAK: a communication act, where spoken language is "moved"
    Slots: *agent*, *object*, *direction*(*from*), *direction*(*to*), *instrument(MOVE)*.
11. ATTEND: paying attention to a communication act. The opposite of SPEAK.
    Slots: *agent*, *object*, *direction*(*from*), *direction*(*to*).
12. DO: is a causal relationships, where a specific agent initiates some other action.
    Slots: *agent*, *object*, *direction*(*from*), *direction*(*to*), *instrument(PTRANS, MOVE)*.

Besides these *primitive acts* the theory also describes the existence of *picture producers*. Basically, these refer to a rather similar idea as that of using frames for representing objects (refer to the previous paragraphs on frame structures). Picture producers describe entities in the real world, and therefore they can be the *agents* in the *primitive acts*.

Having defined constructs that can represent entities and possible actions, there is one more matter to be dealt with in order to be able to represent a world and reason in it, and that is a construct that helps reasoning about existence in a specific world. In Conceptual Dependency theory such a construct consists of *state* descriptions. *State descriptions* are manipulated by the direction slots of several of the primitive acts (cf. MOVE, PTRANS, EXPEL, INGEST). Since primitive acts can be acts that manipulate physical entities as well as abstract entities, additionally the concept of MLOCs is introduced in the theory. MLOC describes a *mental location*, which means that ideas can be transferred from one "mental location" to another.

Difficulties with the semantic level

Whereas the previous phases and the tasks that have to be performed are relatively well understood and clear defined, this is not true for the semantic level. Several difficulties with the semantic level were discussed in the previous section. Among the difficulties discussed where the fact that several representation schemes had very little semantic content, whereas other approaches had a very narrow scope and could hardly be put to use in a more general domain. An important issue is also that growing semantic representation automatically means the increase of representational complexity. Depending on the application of the extracted information this might cause serious problems in applying the knowledge.

Another serious difficulty with representing semantics is caused by the fact that semantics take into consideration the notion of context. At the semantic level, the context has a rather narrow scope, including just a sentence and (possibly) its immediate surroundings (previous and following sentence). Of course such a narrow scope does not suffice under all circumstances, and often there is a wider scope needed for capturing the right meaning of a word. Such a wider scope might capture a complete narrative or a collection of texts or audio streams.

1.11.6 Discourse level

As (correctly) identified by (Allen, 1994)[13] there are a variety of contexts that play a role in language understanding. Several interrelations between these contexts exist. This leads to the effect that on levels that abstract from the syntactic level it becomes harder to proceed according to an ever-increasing representation of the input texts. The representation is sometimes impossible without more discourse knowledge, and discourse and common sense knowledge is not generated before the lower semantic levels are taken care of. Lets consider a classification into contexts as provided by (Allen, 1994) who classifies contexts as follow:

[13] Allen (1994) provides an introduction in *Discourse* that forms the basis for this section. Here we will use a few of the examples provided in that book to help describe the view on Natural Language Processing we propagate here. The interested reader might therefore refer to the respective chapters for more information on discourse structures.

	Specific/Local Context	**General/Global Context**
Situational Context	Bob is the actor. A Honda is acted upon. Its speed is 65 mph on an asphalt road. It is a windy road. The next bend in the road is just visible….	It is 1989. The location is south England. Motorcycles typically come with only two wheels. Two wheels are inherently instable. Gyroscopic forces help out. The next gas station lies 35 mls ahead…. MCs can drive 150 mls on a full tank.
Discourse Context	The previous sentence is parsed as (S(NP It)(VP is (NP a windy road))). It is referring to (NP an asphalt road) in the previous sentence. The Honda rides 65 mph.	The bike has been driving for two hours. Its tank has been topped up two hours ago. The windy road was distracting the rider. Now he just realised what was going on.

Table 3: Categorisation of Context (cf. Allen, 1994).

Taking Table 3 as reference, we can now make a difference between the *discourse* level and the *pragmatic (*also *common sense* or *situational)* level.

Where the previously discussed *semantic level* introduces a first semantic analysis (mainly on the sentence level), the level of *discourse* takes into consideration the context on a larger scale, mainly based on the experiences and evidence from the whole narrative. This means that for the store in the table, there is knowledge about the previous whereabouts of the motorcycle rider (as long as it is discussed in the narrative) as well as knowledge about the fact that the rider has filled up the tank (a PTRANS relation in Conceptual Dependency notation), was (is?) distracted by the windiness of the road and started to realise something was about to go wrong (an MBUILD act!).

The course of the *discourse* does NOT tell us anything about the general world facts, neither does it tell us that something is going wrong, or give us a clue what is about to go wrong. We can only reason upon the facts that are available in the current discourse that builds upon the narrative(s) under consideration so far.

Consider the last sentence in the global discourse context: “Now he just realised what was going on.” The meaning of such sentences is hard to formalise, and even if an appropriate formalism can be defined the question remains the same: how to interpret this sentence in the light of the facts?

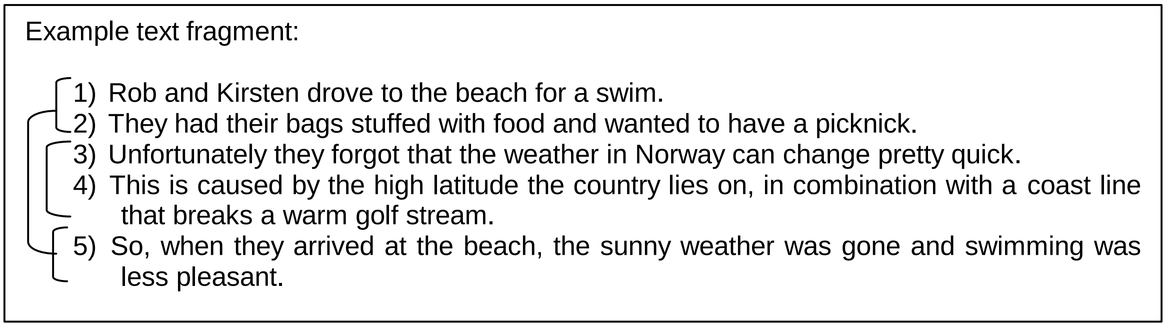
Clearly the strive for “objectivity” in formal approaches is not always compatible with subjective contexts. Formalisms that are theoretically able to represent such sentences on the semantic level as well as on the discourse level (and the situation, pragmatic level?) produce awkwardly complex descriptions[14]. Again, even having a convenient representation formalism (and the belonging parsing and inferencing technologies) do not guarantee that a text is understood. This is because important world knowledge often lacks in language processing approaches.

[14] This effect can already be seen when analysing representations of semantics using Conceptual Dependency theory. It is quite possible that there are graphs needed that span over 5 pages, in order to represent one paragraph of text.

Cue phrases for Structure	**Typical Use**	**Cue phrases for Semantic Relations**	**Typical Use**
by the way	*start digression*	and	*continuation*
anyway	*end digression*	because	*causation/reason*
bye	*end dialogue*	but	*contrast*
first	*start enumeration*	furthermore	*new subtopic*
next	*new item*	however	*contrast*
last	*new item*	meanwhile	*new topic (simultaneous)*
incidentally	*start digression*	so	*conclusion*
now	*introduction new topic*	then	*causal/temporal*
OK (right)	*close topic*	therefore	*summary*

Table 4: Example cue phrases (cf. Allen, 1994).

Another important topic in discourse analysis is the topic of *discourse segmentation*. In many cases, a given discourse is interrupted and a different discourse is started. Often there is a return to the original discourse after the intermediate discourse is closed. Especially German is known for its, often lengthy, sentences that can describe a variety of discourses in one sentence. In such cases special discourse strategies have to be applied in order to keep track of the ideas that lie behind the discourse(s) at hand. Examples can also be found in English, where the difficulty of discourse segmentation is often found on the supra-sentential level. Discourse segmentation can be identified by so-called cue phrases, which can be used in identifying chances in discourse focus based on typical cues like enumerations, but also through more explicit utterances like first .. next .. last, etc.



Example text fragment:

1) Rob and Kirsten drove to the beach for a swim.
2) They had their bags stuffed with food and wanted to have a picknick.
3) Unfortunately they forgot that the weather in Norway can change pretty quick.
4) This is caused by the high latitude the country lies on, in combination with a coast line that breaks a warm golf stream.
5) So, when they arrived at the beach, the sunny weather was gone and swimming was less pleasant.

Figure 17: Building Discourse through Text Segmenting

Knowledge like that represented in Table 4 can now be used for segmentation of a discourse in to its “sub-“discourses. In relation to this we can re-introduce the topic of pronoun resolution, but now not on sentence level, but on the supra-sentential level. Pronouns can refer back to other discourse segments that are followed by different discourses afterwards. In such cases it is nearly impossible to refer to such discourses by using *history lists* and *tense analysis* only.

History lists are lists of nouns, noun phrases (or picture producers, cf. (Schank 1977)) that are stacked in their order of appearance. At the time that a new pronoun is found in the text, it can be resolved by taking the last item from the stack that satisfies the constraints on tenses and gender. Such approach, although simple in nature, can help to a certain level, but certainly are limited in use. From the conversation in Figure 17, the initial discourse started in sentence 1 and 2 is disrupted in sentence 3 and 4, and finally continued in sentence 5. An approach based on history lists would not be able to resolve the pronoun “they” in sentence 5, while its discourse is disrupted by sentence 3 and 4, and therefore the stack will have noun phrases stacked that belong to the other discourse.

A possible solution for resolving discourse can be found in techniques that build discourse segment hierarchies. Building a hierarchy with sentence “blocks” naturally require the analysis technique to be able to deal with the identification of such segments first. More intelligent approaches make use of “triggers” in the text that can identify the start and end of discourse segments (cf. the cues in Table 2). There is evidence that using *cue phrases* for discourse segmenting works significantly better in spoken language than in written language. Many typical “spoken” utterances do not usually appear in written text (cf. “OK”, “by the way”, etc.), whereas they are clear identifiers of a break in current discourse, or of the end of a certain communication.

Nevertheless, it is quite possible to identify discourse context in written texts as well, and particularities in text structures might play a role in that. So is it often possible to identify a certain group of listed items as belonging to a specific discourse context, and thereby resolving pronouns and cross-references. Consider the following text fragment (free after Wiseman, 1993):

Local conditions will determine the type of shelter you build. In some cases you might want to put up a tent. However, do not put your tent:

1. *on an exposed hill-top,*
2. *on valley bottoms and deep hollows. It will be more liable to damp and frost at night,*
3. *near solitary trees, where it attracts lightning,*
4. *near bees’ or hornets’ nests……*

A wrecked plane or vehicle might also provide shelter…….

It is clear that the enumeration in is a list that contains references that can be traced back to the *tent* concept that was the central focus immediately before the enumeration. Several approaches exist that perform such kind of analysis, where a central focus is maintained and references to it are either based on grammatical knowledge, or sometimes even on probabilistic knowledge about which of the currently maintained discourses is most likely to be the right one. Other approaches analyse causal or temporal relationships between sentences and entities described in them. A good study on different approaches for anaphora resolution and the types of models that these approaches produce can be found in (Hirst, 1981 and 1987). Last but not least, there are also several approaches to be found that focus on the definition of grammars for discourse description (Rumelhart, 1975).

1.11.7 Pragmatic level

At the pragmatic level it is tried to resolve all utterances, ambiguities and utilise as much world knowledge as necessary (available?) for doing so. Complete pragmatic, semantic content analysis is regarded AI-complete[15], and it is therefore not expected that a complete solution will be available on short term. Much more likely is it that there will be domain/application specific solutions that will work better or worse under particular circumstances.

Understanding semantic representations. What does a semantic representation represent?

The first often-encountered question is about the *understandability* of semantic representations. Naturally this is very much depending on the type of representation that is chosen, but there are some observations to be made that are more or less generally valid.
Basically there are two dimensions on which one can discuss these observations. On the one hand there are the five above-mentioned tasks in which semantic representations play a role and on the other hand there are the several types of representations that can be used.

REPRESENTATION: TASK:	Semantic Networks	Connectionist Models	Statistical Models	Ontologies/ Hierarchies
Structure generation	++	++	++	++
Visualise/making Explicit the Structure*	++	- -	+??	++
Context based translation	++	++	+??	+??
Query Formulation and Answering*	++	+-	++	++
Content based Search*	++	+-	+-	++

Table 1: Semantic representations and the tasks they can be used for. The '*' marks those tasks where the "human understanding" of the representation structure might be important. (where ++ means "applicable", and – means "not possible").

Depending on the application scenario of extracted information, several representations are more or less useful. Dimensions that are not taken into consideration in the evaluation of table 1 are complexity (and thus speed) and error proneness. A few examples:

[15] By analogy to NP-completeness in complexity theory, "AI-complete" is a term, first mentioned by Fanya S. Montalvo, to indicate that the difficulty of a computational problem is equivalent to solving the central AI problem, i.e. making computers as intelligent as people (Mallery, 1988).

- Visualising semantic networks can be done using a 3D representation of a node structure (for example: imagine a universe with planets that are grouped in solar systems), or using simple graphs (f.e. Self Organising Graphs (Eades, 1984)).

- Visualising a hierarchy using “hyperbolic views” (original algorithm (Lamping, 1996); applications of it in a.o. OntoBroker, (Decker, 1999)), or simple graph like representations for a user to browse. In such cases relations that are not easily seen otherwise (cf. Transitive relations) might become obvious at a glance.

- Using Statistical methods for Query Formulation and Answering. In such scenario’s queries are formulated using knowledge about word frequencies and probabilities about co-appearance of words in a representative corpus of a particular language. When such corpus knowledge is available query formulation can be supported by offering wordlists with related words on the fly, or using the knowledge about the corpus to deal with “term mismatch” in the query while performing automated query expansion.

- Content based search using neural networks is usually not of very high quality, since the search basically breaks down to pattern-matching on documents. Although rather fast in many approaches, the lack of information about the contents of the patterns often leads to poorer results as in approaches that do represent information about the artefacts that play a role and their interrelationships.

It is generally regarded possible to combine approaches in order to resolve deficits in single approaches. Not many practical applications of hybrid approaches are known, which possibly has to do with the combined complexity of the single solutions.

A consideration worth to be repeated regarding the usage of NLP for document retrieval is provided by (Croft, 1987):

- “Document retrieval is a different task than, for example, story understanding and does not require a complete, unambiguous interpretation of the text passages involved. There is [some] evidence that even simple NLP techniques can provide useful information for document retrieval systems.

- The importance of an individual request and the user’s interpretation of meaning of that request provide an inherent limitation on the NLP involved (Sparck Jones, 1984). Rather than attempting to process all document texts independent of individual requests, the NLP component of a document retrieval system should be used to analyse a request plus the texts of only those documents that potentially address the particular information needs expressed in that request. Additionally, while a traditional NLP system must have all the needed domain and linguistic knowledge specified in advance, a document retrieval system can acquire some of this knowledge as needed from a human expert during a search session (Croft, 1987).”

Although many new approaches appeared since 1987, the general underlying ideas mentioned here still hold. Even though approaches become more sophisticated, it is still not true that document retrieval is feasible when using a complete, deep and error free parsing of natural language for representation of meaning and semantics. Therefore, it is still necessary to compromise and find the best possible representation that delivers high quality retrieval while not being too complex. Where academic research is often focussed on the quality and completeness of such representations, the industrial research institutes are more concerned with the pragmatic trade-offs to be made. It is therefore not surprising that (as will be shown in the evaluation of state-of-the-art systems later in this paper) the complete range from non-informed pattern-matching techniques towards informed, semantically rich information representation approaches are used for the task of document retrieval.

How good can automatically generated semantic representations be?

Generally speaking, semantic representations that are 100% accurate with respect to both document contents and references to "the real world" are a *contradictio i terminus*. Where semantics is defined as the "meaning of words" it is exactly this meaning that differs from setting to setting and context to context. Therefore it is arguably wrong to search for a semantic representation of a world state that is the same for everybody and under all circumstances. For working against and with large libraries or repositories of documents, it is clear that there is a model imaginable that represents the knowledge and semantic within that set in some way. However, depending on a certain application type or even a specific context, such semantic representations may have to be different. According to how well adapted a representation is to its application area, it will show more or less serious "errors". This phenomenon is also known from the field of Artificial Intelligence, where building expert systems requires a representation of a particular domain in order to be able to perform expert tasks on that domain. For both domain modelling and the automated representation of semantic knowledge present in a document set, it holds that outside of the analysed areas there is a poor performance to be expected. This argument basically holds for all approaches where a fixed model (learned or not) is used for reasoning, and reasoning quality outside of an analysed area cannot be guaranteed. One solution to that problem might be to describe all knowledge in the whole world (cf. CYC: Lenat, 1995). However, besides of problems with the definition of "all" knowledge there will always be problems with the representation that is chosen when using the CYC knowledge base in different settings for different tasks.
Alternatively one might consider building a system that can model on the fly those areas that are of interest (possibly using some documentation and/or user input as basic assumption for generating "world representation").

So, besides the above mentioned more philosophical considerations, it should be said that the quality of semantic representations depends very much on the amount of knowledge that is added in the primary stages of the analysis process (i.e. the previously discussed natural language parsing stages). Depending upon knowledge about the word types (the more word types are known/assigned, the more potential representations are possible), the captured grammar constructions (which are the verb phrases, where are the related noun phrases, what are the agents, what are the objects that are acted upon, which adjectives describe what nouns, etc.) there is a potential to capture more or less of the actual world state seen from the perspective of the current text(s). Another factor that is often neglected in computer linguistic approaches is the supra-sentence analysis where knowledge of several sentences is used in a parallel manner, as to capture all available knowledge plus the relations between concepts in all directions. Most approaches that can be found work in a serial, uni-directional fashion, whereas it is often required to use evidence in all possible directions in order to establish relations and connections between concepts and words. This discussion will be held in more detail in a later section of this paper.

Complexity and Pragmatics in semantic representations

From the above it follows that complexity of semantic representations are not primarily and necessarily depending on the extension of the domain to be modelled, but much more on the richness of the representation that is required. Richness in semantic representations that are automatically generated is primarily defined by :

- The information that the parser is able to generate when parsing sentences,

- The complexity of the representation language/format, i.e. which types of knowledge can actually be represented. (f.e. in many applications it is of utmost importance that besides information on relational knowledge about concepts there is also information available about the frequencies of single words in a specific language.)

In many academic approaches the richest representations are proceeded, and often there are very well defined, high quality solutions developed for the problem of representing semantics on different levels of granularity (cf. (Quillian, 1968) for an first approach to semantic networks and meaning, (Miller, 1995) for a lexical database for English that represents semantic relations between words and word groups, (Hirst, 1987) for an approach that combines selection restrictions with semantic closeness measures delivering a frame-based representation, (Bobrow, 1980) for an ATN based grammar using a semantic parser to traverse arcs and choose among alternatives). Unfortunately, no approaches are applicable enough to be used on such document sets as available nowadays (i.e. in the form of the internet).

On the other hand, many practical approaches that are known from industrial projects are lacking a high accuracy and good representation mechanism for semantic knowledge that is both rich in representation possibilities and practical in daily use on realistic domains and document sets. (Un)fortunately it is still the case that working with realistic document databases requires much human interpretation and intervention, since the knowledge about the world needed to deal with all chaos, noise and lacking information bits is not yet to be solved by automata.

In the section on evaluation of academic and industrial research we will return to the question what the state-of-the-art today actually is, and if or how we can set a direction for the development of such technology within the OntoKnowledge project.

2 Approaches and Applications of Information Extraction

2.1 Application Tasks that utilise Extracted Information

Information Extraction approaches (IE) represents a group of techniques for extracting information from documents, ultimately delivering a semantic “meaning” representation of it. Now, the availability of a semantic representation of meaning enables a large variety of application scenarios. Many technical products like cars (speech recognition), mobile phones, web crawlers, telephone services and service boots in all kinds of environments are using natural (spoken) language recognition and text understanding technology. In practice, most of these application scenarios fall into one or several of the following classes:

1. **Abstracting and Summarising:** aims at delivering shorter, informative representations of larger (sets of) documents.

2. **Visualisation** documents can often be visualised according to the concepts and relationships that play a role. Visualisation can be either in an introspective manner, or using some reference model/view on a specific topic.

3. **Comparison and Search:** find pieces of semantically similar pieces of information. Comparisons can be based on knowledge about a text generated on one of the specific processing levels discussed above (as in IR).

4. **Indexing and classification** of (partial) texts, usually according to certain categories (as in IR).

5. **Translation:** Context driven translation of texts from one language into another. Language translation has proven to be highly context specific, even among closely related languages. Some kind of semantic representation of meaning is needed in order to be able to make good translations.

6. **Question formulation and query answering** in human-computer interaction systems.

7. **Extraction of information**. The exact definition of the difference between information and knowledge varies among approaches. However, it is arguable that *information extraction* refers to the generation of all additional knowledge that is not explicit in the original text. This information can than be more or less elaborate. *Knowledge extraction* would then refer to the processes that follow on the “syntactical level”, i.e. those processes that aim at induction and deduction of semantic, discourse and common sense reasoning.

8. **Induction/deduction of knowledge** based on extracted information. Many approaches from the field of Machine Learning can play a role here.

9. **Task Definition**: also based on the extracted information. Scenarios: robots or electronic services that get their orders through some natural language interface. Goal definition, planning and task execution are then based upon the extracted information (which can be at all discussed abstraction levels).

10. **Knowledge base generation**. Information that is extracted, deduced or induced can be used in other scenarios as well. A “*Knowledge Base*” can be regarded as a typical

"container" for transfer or sharing such knowledge across applications and time.

These ten application types cover close to all of the scenarios that are observed in the analysed approaches (see section 2.2), applications can be classified as belonging to a single class or a combination of them. Another observation we made is that academic approaches tend to focus on the natural language analysis process (cf. Figure 11), typically specialising the process for a single application type. Industrial applications, on the contrary, are often focussed on offering solutions for a variety of business problems. This leads to the development of approaches that combine general functionality with practical, often acceptance-related, considerations like simplicity, speed and accuracy. There is no need to say that especially the trade-off between speed and accuracy of systems has a high correlation with the acceptance of systems by final system users.

2.2 Evaluating current state-of-the-art applications

For a discussion on the state of the art in Information Extraction (with side-steps to information retrieval where ever feasible), it is important to define classification parameters in which to describe such approaches. The current chapter aims at the description of the state of the art behind approaches. Where appropriate, (cited) examples of existing approaches will be used for clarification or comparison of the applicability of certain concepts and ideas in a specific context. A further restriction will be on the level of detail concerning the commercial systems and the more research oriented approaches that are analysed in a study that preceded the writing of this report. Although a large variety of systems is found, typically only the research oriented (often academic) approaches are described to a level of detail that allows quality comparisons on such diverse issues as parser richness, representation formalisms, needed background knowledge, and the required pre-processing of texts that is needed by the several approaches. Consequently, some of the conclusions on systems that are commercially available has to be deduced from circumstantial evidence[16] and is based on either the documentation provided by the respective enterprises, on knowledge gathered from 'hands-on experience' with the approaches, sometimes from direct discussions with the several suppliers of tools that implement specific approaches.

Figure 18 illustrates the classification scheme that is taken as framework for the discussion on the state of the art for information extraction techniques and their embedding applications or systems. The figure visualises the extraction of information from a single (source) text ***i*** (cf. Figure 11) as a *shaded arrow* from the top left to the top right. Based on this elaborated text a variety of applications can deliver different results (cf. the circular arrow on the top right of the figure). Summarising that text, deduction of knowledge, generation of information (meta data) are possible examples thereof. The extracted information can also be used for relating the current analysed text to one or more different texts (cf. document stack at the bottom left of the figure). These texts are analysed (hence the shaded arrow from the bottom left to the bottom right of the figure, labelled text ***k***). The results of that process are in some way compared to those of the original (source) text. The outcome of such a comparison process can be a document ranking or the generation of advice on the relevance of a particular document or text. Such generated knowledge can form the basis for a push component pushing the respective documents to a particular space (f.e. the *inbox* of a particular user, a specific news group, a ticker tape on a news service, etc.). The user in the middle left is the one that originally originated the process by providing a source or input text (upper left) and, in some cases, a document set as targets for analysis. Both the former and the latter documents can ideally stem from a variety of sources that can be produced in rather diverse manners (speech recognition, emails, text files, etc.).

During the discussion of approaches, the graphic of Figure 18 is used to show on which levels and

[16] Business practices does not often allow for deep comparisons. Attempts for deep comparisons are made nevertheless, see i.e. (Tegenbos, 1997).

which scenarios a particular application focuses. Additionally, we will discuss the requirements on prior knowledge that the different information extraction approaches have. This is because there are severe limits on the richness of the extracted information from texts that is defined by the semantic knowledge available for use at the information extraction phase.

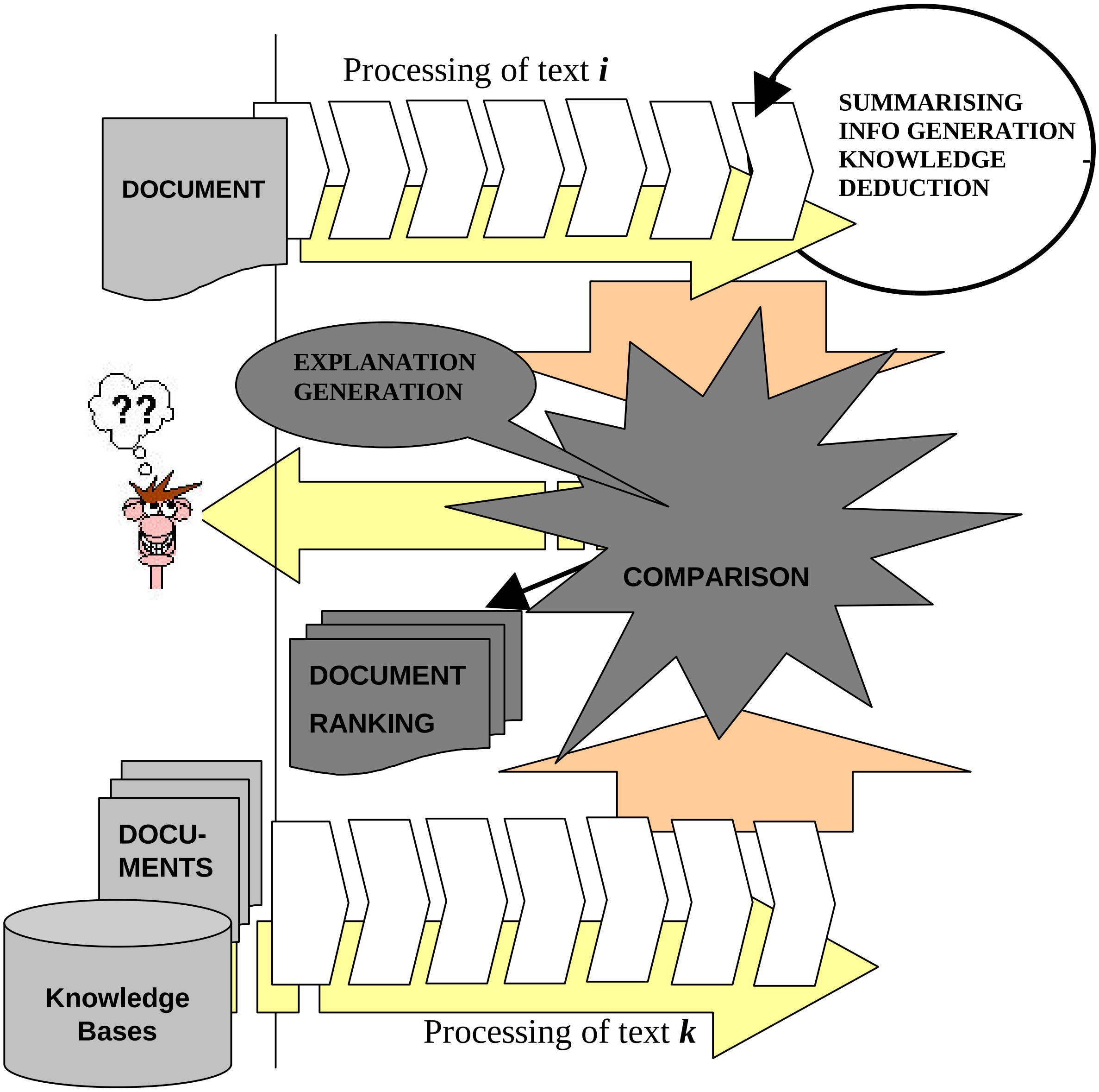


Figure 18: Applications of Information Extraction techniques put into perspective.

Prior knowledge can be of *syntactic* or *semantic* nature (or a combination of both). The former refers to knowledge on morphology, word frequencies in a certain language, or other meaning independent information. The latter refers to techniques relating words based on similarity of meaning (f.e. WordNet: Miller, 1995), relates words according to their meaning in a certain context, describes information as spatial or temporal relations etc. In the general case of information extraction, some knowledge has to be available *a priori.* Without some kind of knowledge modelled before hand, the possibility to be able to extract any useful information from some arbitrary source is questionable. There seems to be consensus on the fact that the more a priori (semantic) knowledge is available, the better the chance to create meaningful and sinful representations of a certain context, situation or circumstances. However, semantically rich representations usually come with severe restrictions on generality of representation (domain independence) as well as on computational efficiency. It is therefore interesting to analyse the solutions that are implemented in some of the representative approaches in the field.

Where a literature study forms the theoretical basis for this state of the art report, a study on approaches available on the worldwide market place has been conducted in order to cover the application side of information extraction. During the last three years in the least (if not longer) numerous suppliers of document management software claim to deal with information extraction in some way or another. Despite decades of research in natural language processing, few applications or products implementing such research are widely known or broadly applied by people not researching or developing in the field. During the last few years there seems to be a renewed interest in the possibilities of natural language, and one reasonable explanation for this seems to be the very simple fact of increasingly faster computers becoming available on the market. The processing of natural language has its requirements on computing power and memory (certainly when using “richer” semantic representations). As discussed in section 2.1, there are many applications that can be “powered” by information extraction techniques as defined in this report, but on the other hand there are as many applications or software solutions that claim more than they deliver in that respect. This goes so far that even relatively standard key word search engines, for marketing purposes, are claimed to perform “knowledge management”. This makes evaluation of such approaches a non-trivial task. However, it seems that the framework proposed before provides enough “grip” to discuss those differences. For the discussion we have chosen to take approaches that share certain characteristics:

- they are *representatives of a force majeur* on the global market,
- the approach is either developed with or for a major player on the global market,
- the approach or application represents a paradigm that is pertinent to the OntoKnowledge project,
- there are good possibilities for an evaluation, i.e. it is possible to get information on the tool or its technology, or there is demonstration software easily available,

A complete list of the approaches under consideration is found in Appendix A:

2.2.1 Inxight

As a Xerox New Enterprise Compnay, Inxight markets a range of products that are initially developed as Xerox PARC inventions. The product that is sold by Inxight consists of mere components that are licensed for inclusion in third party software. The core component is a linguistic component that allows for the linguistic analysis of texts. Furthermore, the product suite includes a summariser, a categorizer and a tool for searching "things", referring to the extraction and analysis of proper nouns in texts.

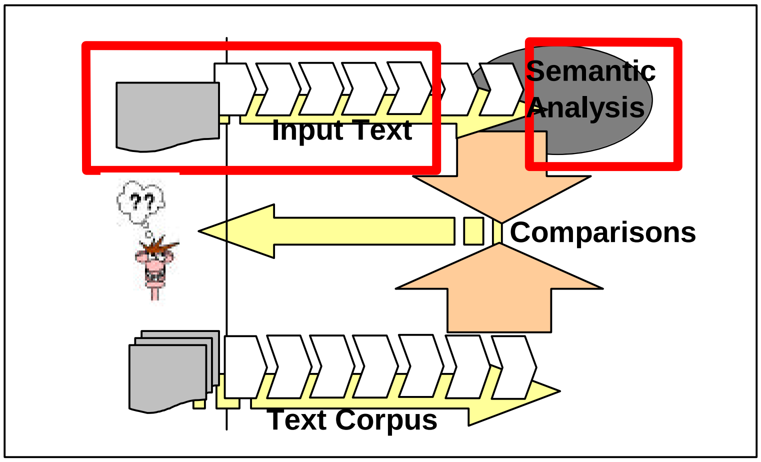


Figure 19: Competence areas of Inxights' component technology (cf. Figure 18)

At the core of the Inxight product range stands the LinguistX platform. The Platform contains a set of components with each their specified function. Basically the LinguistX platform follows the outline of classical natural language processing outlined in Figure 9, applied to single documents. Figure 19 shows this competence area of Inxight, and its focus on NL analysis in a traditional sense. This means that the core consists of the following components:

Lexical Level (cf. Figure 9): *Tokenisation* and *Part-of-Speech* tagging. Traditionally, tokenisation takes care of the identification of word boundaries, abbreviations, etc. In the Inxight tool an analysis is performed at this level for identification of contradictions.
The part-of-speech tagger uses standard grammar approaches for identifying word types and the roles they play within the sentences. Given this functionality, it is highly likely that a "constituency grammars" is implemented.
At the same time there is an attempt to identify the language in which a particular document is written. This is done using the obvious identification of language-specific characters that are used, but is also based on word frequencies, which form the basic fingerprint of a particular language.

Morphological Level: The Inxight Platform provides tools that perform *stemming*, *inflection* and, where needed, *compound word* analysis.
The stemming component can identify root forms of words, thereby using a lexicon for generation of the complete dictionary citation word forms. For purposes of query extension, one can either perform a stemming on all words, or take the opposite approach: inflection. Inflection techniques produce all possible word forms from a given root.
Compound Word analysis is needed in languages in which single nouns can be combined into new words. Examples of these are given above, languages that are especially known for having compound words are Dutch, German, Finnish and Japanese.

Syntactic Level: The Inxight LinguistiX Platform provides *noun phrase* extraction techniques at the syntactic level. Here the opposite is done from Compound word analysis, in the sense that nouns that co-appear in specific phrases are analysed as being highly related. Such knowledge can be used for a more precise representation of specific documents.

Summaries: The Inxight Summariser has been build on a wrapper technology that extracts specific features from documents in order to determine its type (i.e. newspaper articles, letters, etc.). Focus of the summary changes according to these types, which are then used for individualised extraction of the most important sentences in a document. Important sentences are identified by using characteristics as their place in a text and the occurrence of specific words with a high information value (either specified as being of interest by a user, or determined by a statistical approach).

Based on the outlined technology, the Inxight product suite delivers support in the following scenarios:

- **Abstracting and Summarising:** The summariser tool takes as input a document and produces a summary as output, either based on statistical knowledge about information value of single words combined with knowledge generated by a wrapper, or based on the wrapper knowledge plus user defined concepts that are of pertinent interest.
- **Indexing and classification**: The software supports the classification of documents. A corpus with preclassified documents is used to “train” the categoriser. This learner typically refers to usage of statistical techniques (most probable), connectionist networks or information theoretical approaches.
- **Extraction of information**. Extraction of information is performed according to the phases described above. The “highest” level of knowledge that can be retrieved using the tool suite is on the syntactic level, where sentence fragments are identified and grammars are applied.
- **Induction/deduction of knowledge**: The LinguistX Platform does not support induction or deduction of explicit knowledge for use in knowledge intensive approaches. However, it does use inductive techniques for learning wrappers.

The tool suite that is provided by Inxight principally is able to deal with the basic computational linguistics approaches. The actual knowledge generating steps that follow on these basic stages are not implemented by Inxight (unless perhaps for some statistical heuristics that are used in summary building and categorisation) .

2.2.2 AskJeeves

The AskJeeves service is a product by Ask, a US based company specialising in offering "humanised" search engine interfaces. The product allows people to ask natural language questions. If the question asked shows similarity to one or more questions in the AskJeeves knowledge base, then these questions are returned to the user. On top of that query strings are generated that point to several search engines (cf. Figure 21). If the user decides to proceed with one of the alternatives offered by AskJeeves, rather than proceeding with his original question, the answers that are available from the AskJeeves knowledge base are loaded and presented to the user.

On the technology site AskJeeves uses a natural language grammar approach, combined with lexicons and information on the scarcity/usefulness of particular words. This is clearly seen in the example of Figure 21, where the noun "motorcycles" is overruled by the adjective "dutch". The adjective "dutch" is resolved as being similar to "Netherlands" and "Holland". Based on this query extension mechanism, a number of matching questions are found in the AskJeeves knowledge base and returned to the user. AskJeeves also reports the use of editorial services for enhancing the result base that is maintained. Therefore, the answers on the AskJeeves predefined questions are of some quality, whereas the answers on the queries sent to the search engines (as proposed on the bottom of the AskJeeves user interface) are of the quality that is known from standard, key word based search engines.

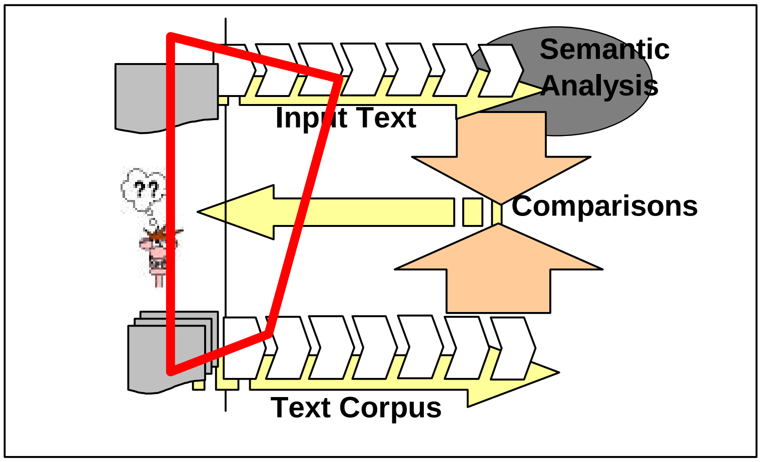


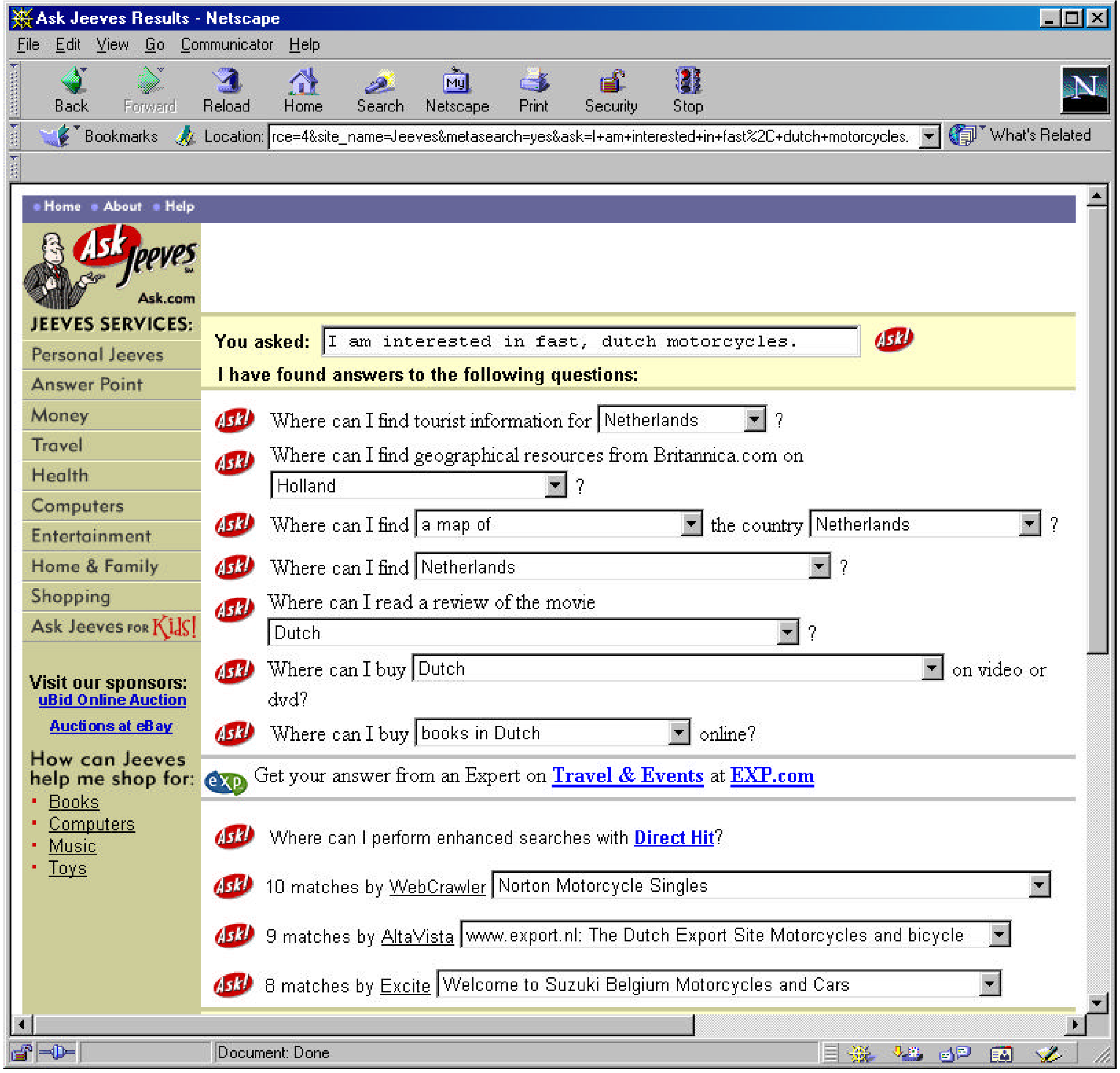
Figure 20: The AskJeeves core competence area (cf. Figure 18).

Figure 20 shows the focus of AskJeeves on the grammatical analysis of input sentences, whereas the retrieval of results is based on a much more shallow process of analysis, often even keyword based only. The AskJeeves engine generally proceeds as follows:

a) first the engine performs *syntactic processing* of the question. Questions sentences are interpreted, part-of-speech tagging is applied and a look up on word frequencies or interestingness (which can be table based) is performed.
b) The words (nouns, proper nouns or adjectives) that constitute interesting words are expanded using a lexicon.
c) The expanded query terms are taken for a question lookup in the AksJeeves knowledge based.
d) The question that is finally selected refers to answers in the AskJeeves knowledge

base (currently reported: 7 million answers).

e) As extra functionality, AskJeeves also functions as a meta-search engine that forwards queries to other major search engines on the web. These answers (of uncontrolled quality) are forwarded to the user.



Location: rce=4&site_name=Jeeves&metasearch=yes&ask=I+am+interested+in+fast%2C+dutch+motorcycles.

You asked: I am interested in fast, dutch motorcycles.

I have found answers to the following questions:

Where can I find tourist information for Netherlands ?

Where can I find geographical resources from Britannica.com on Holland ?

Where can I find a map of the country Netherlands ?

Where can I find Netherlands ?

Where can I read a review of the movie Dutch ?

Where can I buy Dutch on video or dvd?

Where can I buy books in Dutch online?

Get your answer from an Expert on Travel & Events at EXP.com

Where can I perform enhanced searches with Direct Hit?

10 matches by WebCrawler Norton Motorcycle Singles

9 matches by AltaVista www.export.nl: The Dutch Export Site Motorcycles and bicycle

8 matches by Excite Welcome to Suzuki Belgium Motorcycles and Cars

Figure 21: AskJeeves output to the statement: "I am interested in fast, dutch motorcycles."

Analysing the statement "*I am interested in fast, dutch motorcycles*." AskJeeves correctly identifies the noun-phrase "*fast, dutch motorcycles*". The noun-phrase contains the adjectives "*fast*" and "*dutch*" and contains the noun "*motorcycles*". Whereas the regular interpretation of such a phrase let the adjective play the role of a descriptor or refinement of the noun, AskJeeves analyses these words as being of equal type. Interestingly enough, AskJeeves regards the adjective "*dutch*" as more important as the noun, which would state the center of our actual interest more closely. Accordingly, the question (templates) that are retrieved from the database are referring to all possible interpretations of the adjective "*dutch*" such as "*Holland*", "*Dutch*", "*Netherlands*" whereas the noun that represents our real interest ("*motorcycle*") is not taken into consideration. Theoretically, this could be the case while the database with questions does not contain references to motorcycles at all. In order to test this, we took away adjectives and ran the query again. This time all five proposed questions referred to motorcycles.

The sole application area of AskJeeves is that of scenarios in which search plays a major role. AskJeeves could be used for user profiling (based on extracted keywords from posed queries), but does not attempt to visualise such knowledge, neither does it perform abstracting or summarising, translations or induction of new knowledge. The knowledge base that AskJeeves maintains contains questions and relatively high quality answers on them. As long as this knowledge base can be targeted, results are available. In case there is no query template which reflects a particular users specific interests, the only way in which to proceed is to query standard search engines. AskJeeves provides an integrated interface to those engines.

2.2.3 Semio Corp.

Besides the scenario where an explicit search for information is performed, natural language processing techniques can also be used for the opposite scenario. In such cases no searching for specific information is performed, the aim is to analyse a set of document and communicate to a user the contents of such a document set. A representative of this kind of applications is SemioMap, which is part of a family of (mainly three) products marketed by the California based Semio Corporation[17]. The complete product suite contains *Semio Builder* for extraction of the initial document models, *Semio Taxanomy*, which automatically builds taxonomies of key concepts based on a set of documents and *SemioMap* which visualises in 3D a set of key concepts with relations that is extracted from the same document set (cf. Figure 24).

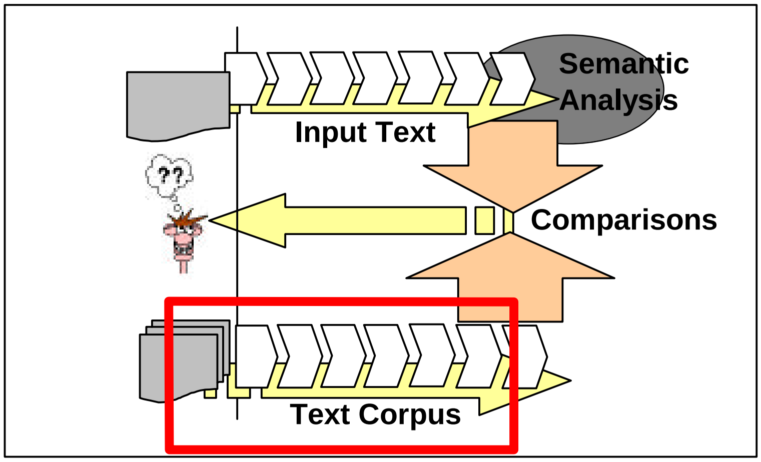


Figure 22: Areas covered by the Semio product suit (cf. Figure 18).

The technology of Semio is aimed at structuring document sets. For creating such structures, SemioMap performs a three-step analysis on text content:

1. *Text collection* from a variety of sources such as intranets, the Internet or server located files.
2. *Phrase extraction* for finding relevant, informative phrases from the text.
3. *Phrases clustering* into a lexical network structure,
4. *Visualisation* using a Java applet viewer (cf. Figure 24),

The basic Semio technology is based on *computational semiotics* (Godlfarb et al., 1997). Computational Semiotics uses as its basic element of “understanding” a construct called *knowledge unit*. A “Knowledge Unit” in semiotic approaches, is a granule of information encoded into a structure (Gudwin, 1999). Knowledge units are used as the atomic elements of intelligent systems. The fundamental claim of *computational semiotics* is that the world is not to be understood in terms of clear, universally valid, objective symbols, but that there is a complex, dynamic world that can only partly (and subjective) be observed. This leads to a taxonomy of types of knowledge in which a world model is described. The interesting fact is that there are *active* and *passive* knowledge types identified. The latter group, the passive knowledge types, is both classified according to its functionality (designative, appraisive or prescriptive) and to its structure (rhematic, dicent).

Rhematic knowledge is the semantic that can be assigned to isolated words in natural language.

[17] The company marketing this software is called Semio Corporation and can be found at http://www.semio.com/.

This type of passive knowledge refers to sensorial experiences, objects and events. Sensorial experiences can be represented e.g. by adjectives, objects by substantives and events by verbs. Rhematic knowledge is said to refer to the semantic memory of an intelligent system.

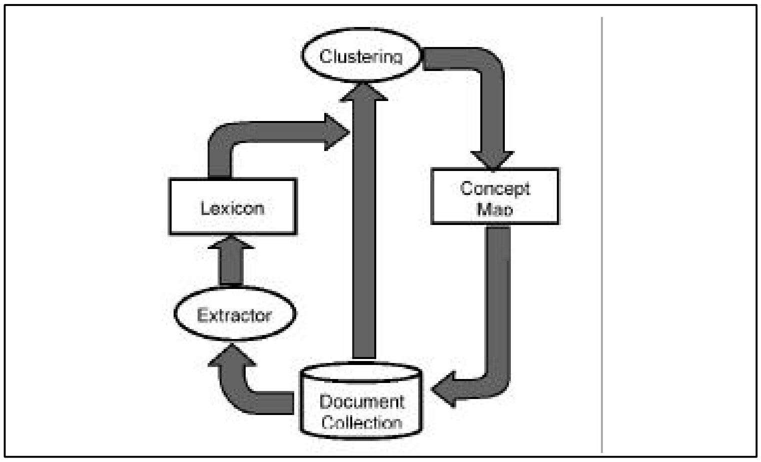


Figure 23: The Semio process

Dicent knowledge describes what is usually found as episodic memory. It deals with propositions and their truth values. Opposite to rhematic knowledge, in which single words are analysed, dicent knowledge refers to sentence fragments or phrases. Truth values for these phrases, for example, can be represented in fuzzy logic.

Next, there is the *apraisive* knowledge. Although it is not completely clear what *apraisive* knowledge refers to from the literature, the most probable interpretation is that verbs with a semantics describing roughly judgements, goals, evaluations of being (like: desire, repulse, fear, anger, hate, love, pleasure, pain, comfort, discomfort, etc.). Such classification of verb in groups sharing a certain type of semantics is also found in (Schank, 1977), where relation types like PTRANS, ATRANS, MTRANS, etc. are defined as a group of *functions*, or *actions* that share a specific semantic meaning.

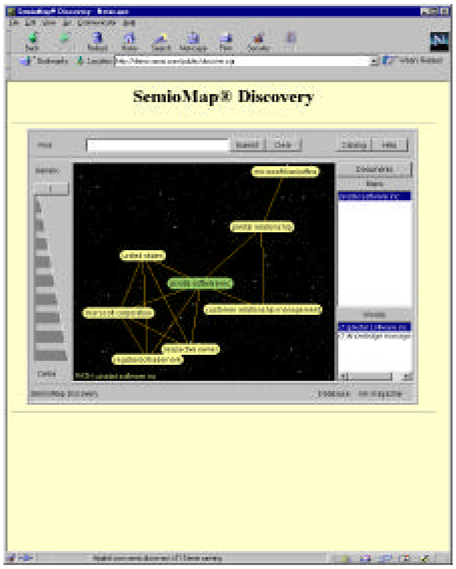


Figure 24: Semio 3D Visualisation Java Applet

Prescriptive knowledge describes functions that act on the world, like the execution of plans and the like.

The active knowledge type mainly includes so-called *argumentative knowledge*. Argumentative knowledge describes reasoning processes. It can be the rules a truth maintenance system uses for deciding on truth-values. As examples of this kind of knowledge one can think of logical reasoning like the *modus ponens* and the *modus tollens*.

Applying the theory of *computation semiotics* leads to the construction of semantic net-like structures, where relations and concepts are typed, thereby allowing for a rich semantic representation. There is an argument for using fuzzy logic like techniques for the maintenance of truth within a representation according to a specific state of the world. Generated structures are visualised using Java Applets (cf. Figure 24). At the same time, the *Semio Topic Library* features a predefined taxonomy of concepts where there is a general consensus on its truth. This taxonomy includes about 2000 categories, ranging from common sense knowledge to industrial topics (An example from Semio Corporation: “SPORTS” has a subtype “SPORTS-EQUIPMENT”, which has as subtype “RACKET” having a subtype “SQUASH-RACKET” in turn). This taxonomy can be used by Semio for further refinement into a taxonomy that reflects a certain world state which is described in a specific set of documents.

The basic type of applications that could be covered by the Semio Product Suite are:

Extraction of information: Information is extracted in the form of concepts and phrases that are put into relation to each other. Both concepts and relations are typed, where the different types represent a specific “semantic”.
Induction/Deduction of Knowledge: Clustering techniques are used to group entities based on retrieved (lexical) information.
Indexing and classification: Based on the clusters that are created plus a predefined taxonomy provided beforehand, documents can automatically be clustered into this taxonomy while simultaneously offering the information that is needed in order to extend and refine this taxonomy.
Knowledge base generation: Principally such knowledge can be used for the creation of knowledge bases that reflect a certain “state of affairs” in a specific domain. Such a knowledge base would then reflect the state of the world at a specific point in time (namely when the world descriptions/texts are analysed), reflecting a specific worldview (the representation space that is defined by the types of semantic knowledge).
Visualisation: Finally the extracted and generated knowledge is visualised as a network structure (cf. Figure 24)

For specific applications, especially those where no explicit information is looked for, Semio is a useful tool for visualising “what is there”. An interesting feature of the Semio product suite is the possibility to automatically analyse texts, extract concepts plus their interrelationships and refine a Semio defined browseable taxonomy with this information.

2.2.4 Cartia

Cartia Inc. is a company that has a slightly different philosophy regarding the retrieval of information. Instead of accepting a natural language query for searching, the Cartia product ThemeScape visualises sets of documents as landscapes. Users are allowed to browse through these landscapes, made up of mountains and valleys. Mountains in the landscape represent concentrations of documents that are related in some way, valleys represent the opposite case i.e. no documents are actually found in that area of the map. Documents are represented by tiny circles, and if interesting, a user may mark them with a flag for later reference (cf. Figure 26).

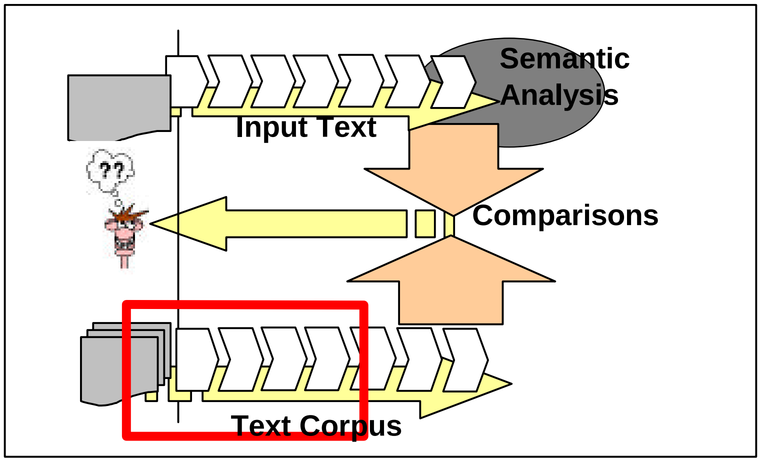


Figure 25: Cartia's areas of competence (cf. Figure 18).

On the technological site, Cartia extracts the most important concepts from a set of documents. A concept vector represents each document in this set. The extraction of concepts (nouns basically) require a linguistic component that is capable of performing a lexical and morphological analysis.
The next step in the process is to cluster the vectors according to their proximity towards each other. Several techniques exist that can perform clustering on sets of multi-dimensional vectors. Typical examples of such clustering techniques are K-means clustering (statistical foundation) or so-called Self Organising Maps (SOM; Kohonen, 1997). Each cluster that is generated represents a number of documents of represented by their vectors. Cartia uses the 5 best concepts of each document to describe a document to the user. On top of that each cluster is automatically labelled with the 2 or 3 most descriptive concepts of the document set in that particular cluster. Cartia also creates a short summary of each document that is stored in a proprietary database for later reference. These summaries are created from several sentences in the document.

The advantage of using techniques like SOMs or K-means clustering is that they are relatively robust against noise in the vector representation. In natural language process this robustness is very useful, since multi-dimensional vectors generated from text documents have a tendency to be noisy due to ambiguities in parsing and interpretation of concepts. The same holds for slang words and spelling errors etc. The map that is finally created in this way provides a means to define which documents fit together, and implicitly also maps closely related (thematically close) concepts. The way in which for example SOM algorithms cluster vectors is as follows:

a) A grid of random points (“cluster centres”, or “seeds”) represented by an n-dimensional vector is generated. The number of points can be fixed or dynamically expanded if necessary, depending on implementation.
b) Input vectors (each representing a document) are introduced into the grid space, and are

classified according to the closest seed. The location (represented by the vector) of this seed and possibly its direct neighbourhood is now changed in the direction of the input vector. Often a decay function is included to make the adaptation strength a function of a seeds' distance to the current input vector.

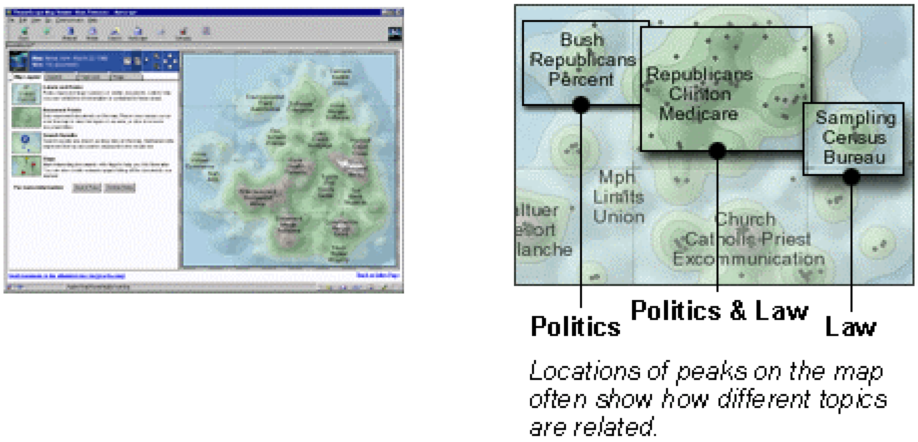


Figure 26: Cartia's ThemeScape visualisation of document sets.

c) After all input vectors are analysed in this manner, certain clusters will be larger and nearer to each other as other clusters. The representation now shows internal relations among documents within a cluster (context representation) and between clusters of documents (thematic "landscape").

d) For each cluster the most frequent concepts can now be taken as a label for that particular cluster. Instead of frequency other heuristics can be used as well. As can be seen in Figure 26, Cartia uses for labelling 2 or 3 concepts taken from the document set that a cluster is made of. The labels that are given in bold black ("Politics", "Law", "Politics & Law") are interpretations on the site of the user, the ThemeScape tool is not actually performing this classifications.

The application types that are covered by Cartia/ThemeScape are the following:

- **Abstracting and Summarising:** ThemeScape abstracts from actual documents by presenting landscape maps of a document set. It provides short summaries of texts for each document that is represented. These summaries are made of parts of the texts (typically the first lines found).
- **Visualisation**: Visualisation of information is appealing, although not always very intuitive. The clustering that is performed is biased by the underlying heuristics of Cartia, which is fixed and therefore shows only one possible view on a document set. This does not necessarily deliver the clustering a user had in mind.
- **Comparison and Search:** Documents are compared to each other and put into relation, rather as being compared to a "source" text or "interest" statement. Therefore, the maps do not represent an answer to a specific query, but rather a possible categorisation of them. Search (on keywords) can be performed within these visualised maps.
- **Indexing and classification**: Classification and indexing is done according to a specific heuristics and leads to an automatically but fixed classification of documents into categories. The classifications are labelled with key concepts that play a major role in the cluster.
- **Translation**: Translations are not supported.
- **Question formulation and query answering**: Cartia/ThemeScape allows a user to use

keyword search as navigation means within a document set. It does not, however, allow information retrieval in the way described in this paper.

- **Extraction of information**: Describing what kind of information extraction is performed by Cartia is a little tricky. An obvious answer would be to state that Cartia does not perform any information extraction, but this is not entirely true. The system extracts key concepts and uses a specific heuristics to cluster these. Both processes create a specific type of knowledge (concepts in the first case, relationships among these concepts in the second case). The systems' functionality is comparable to systems as WEBSOM (Honkela, 1996).
- **Induction/deduction of knowledge**: No knowledge is deduced, or explicitly represented other than the cluster locations and their distances towards each other.
- **Task Definition:** Not supported.
- **Knowledge base generation:** The knowledge that is generated consists of clustered documents and the interrelationships between those clusters. On top of that the knowledge base is filled with summaries of the several documents.

As such, the Cartia/ThemeScape approach represents a group of information representation techniques that visualise knowledge independently from a specific user interest or focus. They are typically used in scenarios in which the main aim is to provide an overview on "what is there". Since this way of representing document sets can be seen as a "compression" of the contents of the single documents, thereby representing some state-of-the-world at a specific moment, the representations can also be used in a timely fashion. In that case, chronologically ordered "maps" can be used to identify changes of interest or focus on specific sites or servers. This can lead to very interesting application areas like the analysis of interest changes in certain areas of research, conferences, businesses, organisations, media and the like. An example of research that performs exactly such a task, and could benefit from the currently described techniques in addition to the techniques they currently use ("factor analysis" and "multi-dimensional scaling") is reported in (Besselaar, 1996).

2.2.5 Verity

Verity Inc. is a USA based company selling information retrieval software. The product suite that verity markets performs a variety of different tasks, including searching, indexing, classifying and summarisation of texts. The architecture of the product suite is component based, such that individual enterprises can choose the functionality that they are looking for.

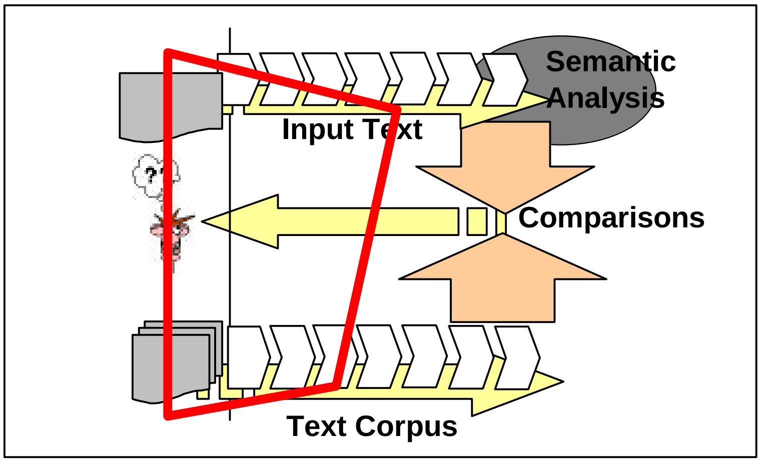


Figure 27: Verity's covering of the Information Extraction arena (cf. Figure 18).

Verity software allows for a variety of classification and indexing approaches to be applied, including standard key word search, rule based document classification (cf. SiteSeer by Aidministrator), full text search and natural language query processing.

Seen from a technical point of view, Verity’s products have several interesting linguistic capabilities. The indexing and search capabilities of Verity are the most relevant for the current discussion. The software is able to:

a) apply *linguistic analysis* to natural language texts. On the lexical and morphological level, Verity software is able to perform *tokenization*, *part-of-speech tagging* and *stemming*.
b) perform *lexicon* look up and use semantic networks (cf. WordNet, Miller 1995) for query expansion.
c) *Clustering*, using user generated rules which, once applied to a specific document, calculates a measurement saying how likely it is that a document deals with a certain topic (classification).
d) perform *information retrieval*. Based on the analysis of queries in natural language or keyword based queries, the software ranks documents according to their interestingness, clusters them or highlights the most interesting key terms.

This core functionality is supported by an architecture that is able to analyse a variety of file servers and the www. A variety of push and pull techniques is used for the distribution of created information. Clustering of Although offering clustering capabilities there are no visualisation techniques available like for example found in Cartia or Semio software. As far as its linguistic capabilities are concerned is the Verity component suite able to perform linguistic analyses of a document in order to retrieve concepts that play a key role in a certain document. The structure of concepts is used for comparing documents on their similarity or proximity of search terms to those found in other documents. The found documents are ranked according to their interestingness, but no explanations on the reasons why they are found are given based on the semantic analysis (this is true for the query-by-example scenario, where a natural language text forms input for a search). When applying the rule-based document retrieval application, there is feedback on which rules contributed the most to the final classification of a document.

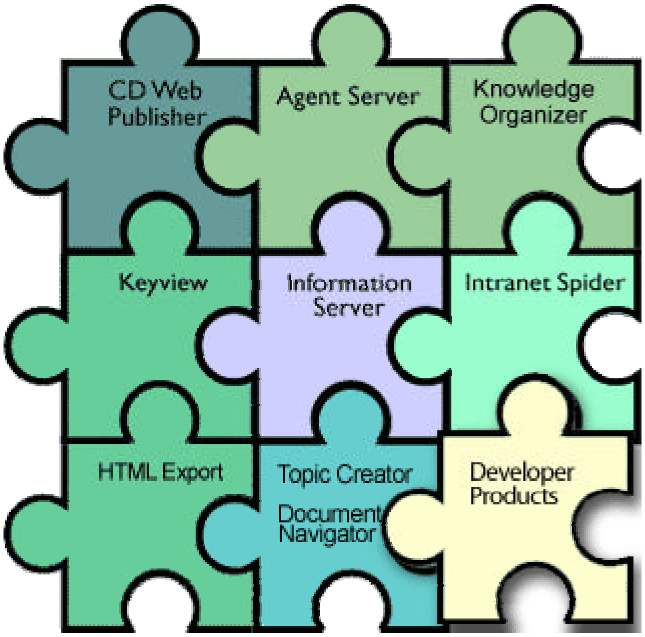


Figure 28: Verity's component architecture.

Figure 28 shows the architecture that lies behind the Verity software. The main components of interest for the current discussion are the "*Information Server*", and the "*Knowledge Organiser*" component. The Verity software maintains for all documents it analysis a word list, default and user-defined metadata (extracted concepts and rule-based categories resp.), and the document's physical file system or URL address (i.e. documents are not mirrored on the local server that Verity is running on).

The applications that are supported by the Verity software suite encompasses:

- **Abstracting and Summarising:** Verity software does not produce document summaries in some sense.
- **Visualisation:** documents are visualised in structures, but not in maps or other graphics or visualisations. Categorising of results according to the Topic Creators' rule base.
- **Comparison and Search:** Comparison and search is performed based on the concept extraction capabilities of the Verity software in combination with Boolean keyword search or rule-based classification. Reportedly fuzzy comparisons can be made between documents, but exact details are not provided. Semantic representation is restricted to word proximity and comparison of concept sets.
- **Indexing and classification**: Verity performs classification and indexing according to the extracted concept sets, or according to created rule bases that describe typical classification classes.
- **Translation**: not available.
- **Question formulation and query answering**: based on the same concept extraction functionality, in combination with thesaurus/lexicon application for query expansion, natural language queries can be analysed.
- **Extraction of information**: The information that is extracted (mainly categories and concepts) can be used for term highlighting in existing documents as a kind of "quick read

guide".

- **Induction/deduction of knowledge**: not supported
- **Task Definition:** not supported.
- **Knowledge base generation:** The Verity software suite maintains a knowledge base with URIs of documents, extracted concepts, automatically created meta data on these documents (i.e. rule based classifications) as well as user defined knowledge.

Verity's information extraction and retrieval capabilities are evaluated as rather basic, whereas the product suites' application software is relatively extensive. Support is provided for a large variety of platforms, file formats and user interfaces. However, no explicit semantic knowledge is extracted and offered to a user, and retrieval of information seems to be on the document level rather than providing pointers to the location of the knowledge in the documents themselves. The software support components show a focus on autonomous support techniques (automated information gathering, push techniques for information distribution, etc.).

2.2.6 Autonomy

Autonomy's core software aims at the retrieval of interesting information, featuring a range of products that utilise this core engine in several distinct manners. The company is a spin-off from NeuroDynamics ADT, a Cambridge UK based company. The core component in Autonomy's software is called *Dynamic Reasoning Engine* (DRE™). This DR-Engine performs four main functions (Autonomy, 1998):

- *Concept Matching*: refers to the task of using an input text as basis for finding similar texts or documents in other sources.
- *Agent Creation*: refers to the possibility of exporting an autonomy internal encoded specification of the structure that is generated by DRE. According to (Tegenbos, 1997), Autonomy is using a set of *extracted keywords* as *interest model*, where keywords are extracted from a text using word frequency lists of the language that is analysed. Information theory (Shannon, 1948) is used for indicating which of the words in an input text are most important.
- *Agent Retraining*: or refinement, refers to the process of using additional texts for refinement of the originally created *interest model*.
- *Standard Text Search*: The DR-Engine analyses natural language queries or just keywords for search.

The technology that is used for information extraction is based on statistics and information theory, whereas the retrieval of similar documents is based on pattern-matching (connectionist) neural networks. Autonomy's DRE technology covers the areas visualised in Figure 29. Information retrieval is based on statistics and information theory, comparison is based on application of connectionist models.

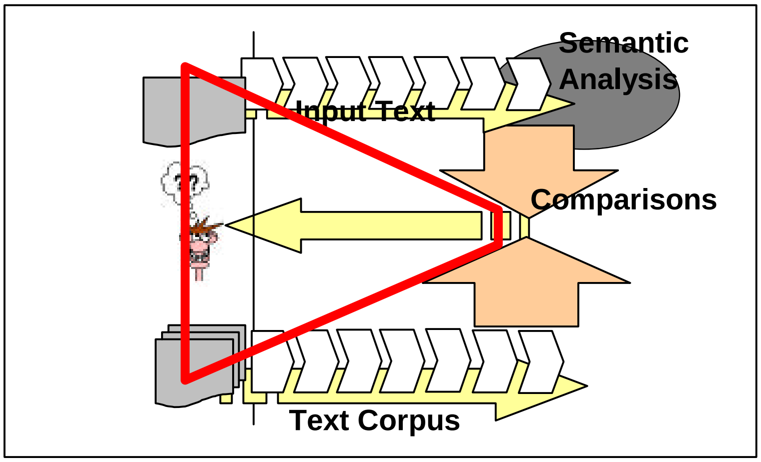


Figure 29: Areas covered by Autonomy's DRE technology (cf. Figure 18).

An analysis by Autonomy takes (presumably) the following steps:

- Phonological level: not applicable, this level is implemented.

- Lexical level:
  - *Tokenisation*: single words are identified.

- Morphological level:
  - *Stemming* (genus): In older evaluations and newer tests using Autonomy there is

no evidence that stemming is performed at all. Adverbs and verb tenses are taken into consideration as written in the original text (cf. Tegenbos, 1997).
  - *affix resolution* and *compound analysis* are not performed by Autonomy.

- Syntactic level:
  - *Pronoun resolution*, *application of grammars* and the *identification of sentence fragments* are not performed by Autonomy. On the contrary, the company makes not having this kind of analyses a major point of critic towards other approaches., the main claim being that “classical natural language processing” has not yet been capable of dealing with the (semantic) ambiguities of natural language utterances.

- Semantic level:
  - There is *no processing at the sentence level*, i.e. no representations based on single sentences are made if a document contains more as one sentence. No alternative search concepts are introduced, which would require semantic knowledge like the knowledge available from WordNet (Miller, 1995).

- Discourse level:
  - Shannon’s Information Theory is used for determination of the information value of single words compared to a frequency wordlist that is available for specific languages and can be “learned” for other languages. Those words with a high information value form the bricks for building an interest model. Autonomy’s DRE engine represents texts as complete narratives by using *statistical approaches* for building and representing dependencies between the search terms. In (Autonomy, 1998) it is suggested that the approach uses Bayesian theory for definition of such relationships. This might refer to the usage of Bayesian networks that can then be used for the definition of a topology for a neural net as well as the initialisation thereof based on the calculated correlations between extracted search concepts.

- Pragmatic or Common Sense Level:
  - No implementation at this level. No “understanding” of ironical or metaphorical utterances are not dealt with by DRE.

Several tests have shown some of the weaknesses of the approach taken, as well as positive points. Autonomy is based on the philosophy that the “meaning” of single words should not be defined using grammars and other techniques from the “Chomskian” paradigm[18], but that this “meaning” of words is solely determined by the context they are used in. Determining the meaning of a word by its context can to some extent be compared to word sense disambiguation in more traditional approaches. In both cases the idea is to disambiguate a particular word either by looking up similarities based on thesaurus knowledge or by comparison with its “context”. In the latter case one can retrieve documents based on similar contexts, even in cases where the specific concept is not explicitly named. In the case of the evaluation by (Tegenbos, 1997), this seams not always to work out in the case of Autonomy’s DR-Engines’ analysis process, and especially this type of dealing with “noise” seems not to work out particularly well.

The basic types of applications that are covered by Autonomy are (cf. section 2.1 for terminology used):

Abstracting and Summarising: DRE can summarise documents and augment them with hyperlinks that point to related texts.
Visualisation: A graphical interface is available that allows a unified view on different types of

[18] Cf. Chomsky (1966) and Chomsky (1968).

documents (emails, reports, news groups, etc.).

Comparison and Search: Autonomy provides *User Profiling, Document Retrieval, Cross Referencing*. All these applications are based on the DRE Concept extraction and comparison engine.

Indexing and classification: *Categorisation* is also provided by Autonomy. In such a scenario a specific agent maintaining an interest model each represent its own category. Documents satisfying the similarity requirements are categorised according to the most similar agent.

Virtues and shortcomings of Autonomy's approach

The merit of Autonomy's approach probably is its departure from traditional linguistic approaches, and thereby implementing an approach that is general and pragmatic enough to perform a relatively good job in real world practical settings. In all its principles the approach is rather unconventional, seen from a computer linguistics point of view, which leads to the effect that the approach has some unique opportunities for dealing with scenario's that are more problematic when using traditional IR approaches. Most prominent of these unique opportunities is its (relative) language independence, due to the fact that no grammatical knowledge is required for analysis.

On the other hand there are some disadvantages with the approach Autonomy takes:

- the narrative to compare to has to have a *certain size* in order for statistical approaches to work, the effects of small "interest statement" in the size of a single sentence can be that irrelevant parts of the "interest statement" are selected for comparison with target documents[19].
- related to that is the fact that there are many mismatches and irrelevant documents found.
- there is *no explicit knowledge* built up. Therefore it is hard to see how Autonomy's DRE can be used for knowledge induction and generation of knowledge bases,

The main problem with Autonomy's lack of "deep" understanding of texts and documents seems to be the fact that especially "single line question" scenarios (cf. AskJeeves) seems to be in need of such "deeper" understanding, since a deeper understanding can be compared to a kind of query extension where available background knowledge is used for annotating a single line query. With only probabilistic and information theoretical information available, an analysis of such a short text will easily by confused by irregular sentences and incidental formulations (cf. Appendix B: test of Autonomy).

[19] Tests showed that in case of small, single sentence questions led to the selection of single words or phrases like "some" or "something like" where selected as pertinent to the stated interest, which in the current case led to the effect that selected documents where also mentioning these utterances, thereby completely missing the context of the question (cf. Appendix B: test of Autonomy).

2.2.7 CORPORUM and Mimir

CORPORUM supports both personal and enterprise wide document and information management - that is management by content. The CORPORUM system is founded on CognIT's Mímír technology developed in Norwegian research labs. This technology focuses on meaningful content rather than odd data or standardised document parameters. CognIT's mission is to capture the content with respect to the interest of the individual rather than address the document itself. There are three essential aspects of this.

- CORPORUM interprets text in the sense that it builds ontologies that reflect world concepts as the user of the system sees and expresses this. This ontology constitutes a model of a person's interest or concern.
- The interest model is applied as a knowledge base in order to determine contextual and thematic correspondence with documents presented before it.

The interest model and the text interpretation process drive an information extraction process that characterises the hit in terms of relevance and in terms of content. This information can be stored in a persistent database for future reference.

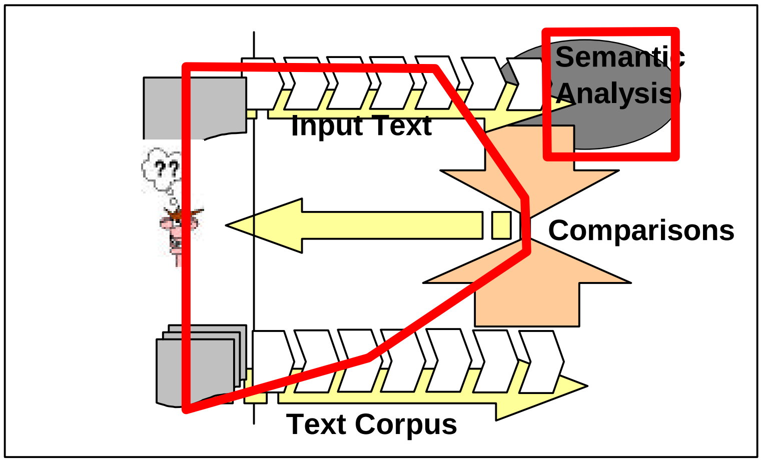


Figure 30: Corporum core competence areas (cf. Figure 18).

The CORPORUM software consists of a linguistic component, taking care of tasks as lexical analysis (tokenisation, part-of-speech tagging and possibly lexicon application), morphological analysis (collation, inflection) and analysis at the syntactical level (grammar application). At the semantic level (cf. Figure 11), CORPORUM performs word sense disambiguation by describing the context in which a particular word is being used. Doing so is naturally closely related to knowledge representation issues. CORPORUM is able to augment "meaning" structures with concepts that are invented from the text. The core of the CORPORUM system (the MiMir engine) is also able to extract the information most pertinent to a specific text for summary creation, extract the so called Core Concept Area from a text and represents results according to ranking which is based on its interestingness towards a specific contextual theme set by a user. On top of that, the CORPORUM system is able to generate explanations, which will allow a user to make an informed guess on which documents to look at and which to ignore. CORPORUM can point to exactly those parts of the targeted documents that are most pertinent to a specific user's interest. Figure 30 shows the competence area's of the CORPORUM software.

2.2.7.1 Overall architecture

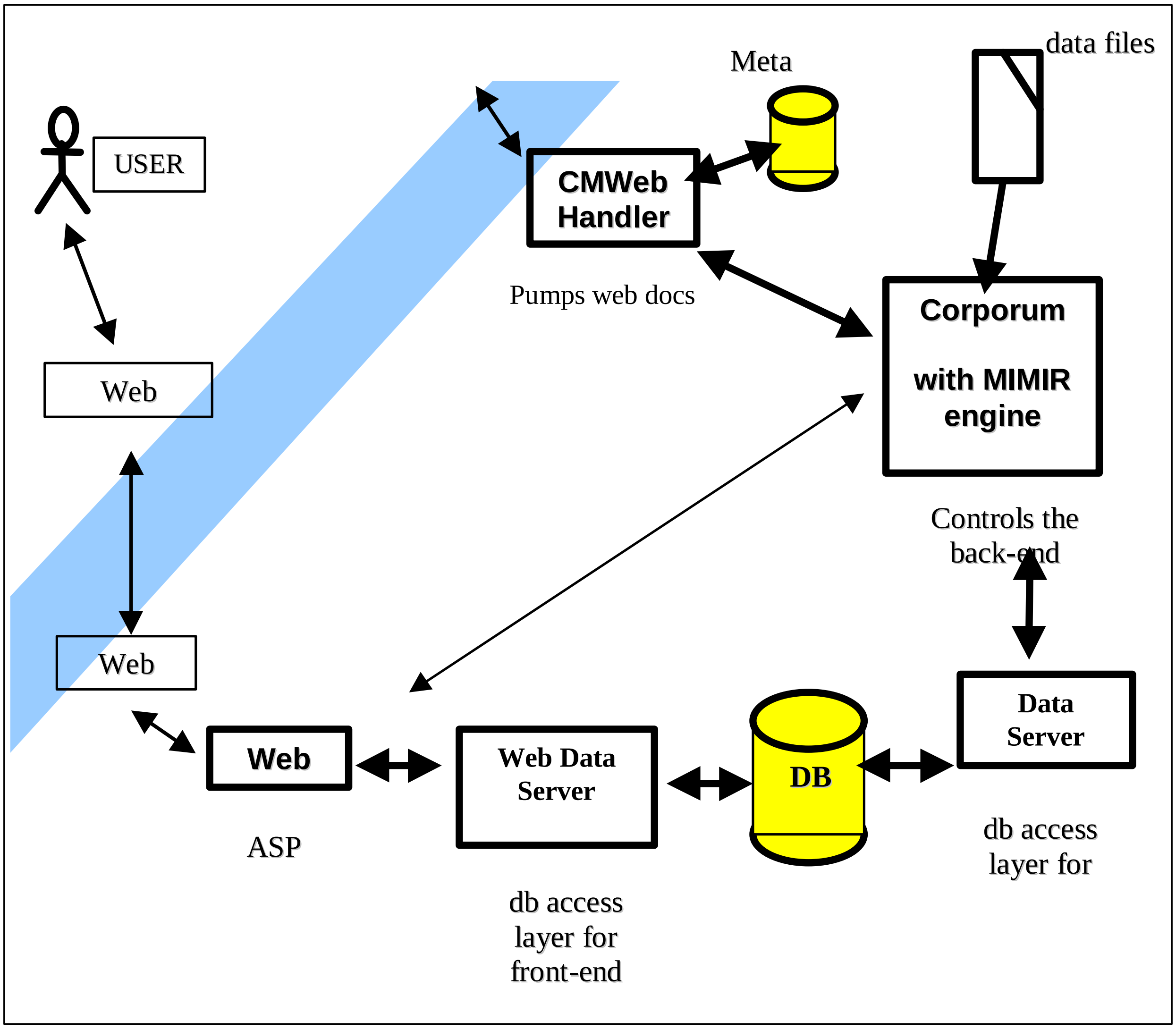


Figure 31: The overall CORPORUM architecture

The overall architecture of the CORPORUM system is depicted in Figure 31. It consists of 4 basic software components each coming with a well-defined API. This implies that CORPORUM can be reconfigured compared to the depiction given here. It also implies that each of the components can be reused in a different setting. This is especially important for the part that contains the MIMIR based engine. This is the heart of the CORPORUM system, but it may operate on an independent basis as long as its API is observed. Because of this it is possible to include a subset of the components shown in various other configurations. One example could be an ERP system that needs contextual indexing of different types of documents.

Figure 31 shows the several components that the architecture consists of. The user may have access to CORPORUM from any web-browser hooked up on the net. Access to CORPORUM is given through a regular ASP functionality that communicates with a Web Data Server. The Web Data Server handles the interface to the database. CORPORUM is suited with a standard relational database. However, any database can be applied. A change to this part will not have any effect on the main architecture since both the Web Data Server and the Data Server are designed according to a pure object-oriented standard. All database calls from CORPORUM components are made on an abstract level. The servers interpret this to SQL-calls or similar of the type required for any given database. The Data Server feeds analysis results from the CORPORUM kernel so that it can be maintained in the database. The kernel consists of several lesser components. The most important of these are the ones that accommodate the MÍMÍR technology. This consists of several algorithms that drive the analysis and information extraction functions. Several of these algorithms apply linguistic rules and information contained in separate files. In order to handle multiple languages CORPORUM will contain several sets of these files. The MÍMÍR technology will be described in more detail below.

The CMWebHandler contains both crawler capabilities as well as document processing functions. CORPORUM can be equipped with a set of such handlers in order to treat different types of document formats beyond standard HTML. The CMWebHandler receives search instructions from the kernel component.

2.2.7.2 Ontologies

CORPORUM applies ontologies in order to establish whether two entities communicate. In order to establish “real communication” both the speaker and the listener must share an ontology. If an author wants a reader to understand what he writes he must attempt to find terms that can serve two basic purposes:

- Express his ideas
- Trigger an understanding of the reader

Only when this is achieved the message can get across. CORPORUM applies these principles in its treatment of the document text to be analysed.

2.2.7.3 The MÍMÍR approach

And Odin hungered for the wisdom that the well could yield. "Nay", said the giant, he stepped up against the mighty god himself. Odin[20] *mustered the other, " You ask a high price". "Wisdom gives power. MÍMÍR has got what you want. An eye and you may drink." Odin ripped his eyeball out and bent over the rim in the shade of the huge tree. He filled his mouth and drank. The well of MÍMÍR was truly strange. As he devoured the liquid that tasted like water he sensed a new divine feeling. Even without his eye he could see so much further.*

-- *Free after Norwegian Mythology*

The CORPORUM products are powered by a technology that was invented by CognIT. It consists of two basic concepts, a concept extraction facility and a resonance algorithm. Concept extraction focuses on the semantics of a text. The approach is rooted in classic information theory dating back to the seminal work of Claude Shannon. However, MÍMÍR goes beyond the mere signal and looks at the concept behind the term. The concept extraction effort combines natural language processing and knowledge intensive methods.

The resonance algorithm enables comparison between the contents of two texts. This implies that the conceptual structures of the two texts are analysed against each other. The resonance metaphor stems from the idea that the match process triggers violent reactions if the frequency of the emitted signal is close to the natural undamped frequency of the objects themselves. In other words if there is a good match then there is high resonance with respect to the content of the text found. The difference in amplitude determines the degree of resonance. In figure 9 this is illustrated. The information model that represents the semantic content of the text that is being analysed will yield

[20] **Odin** - The Norse god of wisdom, war, art, culture, and the dead and the supreme deity and creator of the cosmos and human beings.

a broad, albeit a minor response if there is a form of resonance. Text B is an example where the whole of the model is stimulated rather than an isolated part. Two discrete bars represent mere lexical responses in Text A. Despite high amplitudes they stand out as very isolated with no intermediate amplitudes. The set of amplitudes generated reflects the degree of match and determines the contextual and thematic aspect of a text.

An inherent feature of the resonance algorithm is that it applies a multiplier effect. This is a typical feature of many complex systems, both socio-economic and natural. Multipliers are immediate feedback mechanisms that enable complex structures to prosper despite the fact that it is often supported with very lean input that fuels its growth. MÍMÍR does the same thing with scarce data. It reapplies it in multiple ways almost simultaneously in order to build up an understanding that can justify a conclusion.

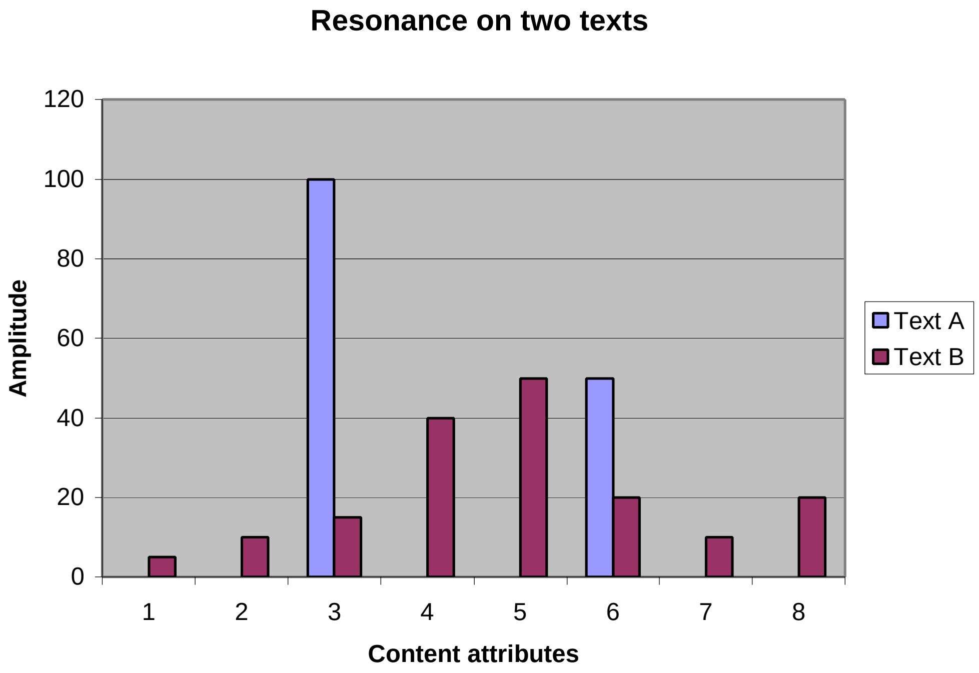


Figure 32: Illustration of resonance. If the match process is unable to find any similarities between the content of two texts there will be no resonance. It may well be that there are a few words that are common to both, but this alone may not yield sufficient response across the full context. Text A shows this type of match result. Text B shows a distributed response across the full text content, thus indicating resonance and a good match in terms of content.

2.2.7.4 Contextual interpretation

The basic text interpretation capability of MÍMÍR is contextual. Framing the context inherent in a document is fundamental if you want to enable prudent indexing and grouping. In such a case you let the content of the document rather than an ad hoc set of keywords determine the indexing. Once you have settled the context it is possible to determine what the document is all about, how it overlaps the content of other documents. Moreover contextual homogeneity assures that quick lexically based searches returns the desired results. The MÍMÍR technology also enables mapping the document content. Since the approach focuses on meaning it is possible to visualize the knowledge inherent in a document. Linked with objective parameters such as name of author and creation date it is possible to build a "who knows what" directory and to measure development in knowledge focus for both individuals and groups.

Example:

ASEAN signs deal to create free investment area

MANILA, Oct 7 - The Association of South East Asian Nations (ASEAN) signed a framework agreement on Wednesday to create an ASEAN Investment Area (AIA) to stimulate investment in the economically battered region. The agreement calls for the creation of a competitive ASEAN Investment Area with a more liberal and transparent investment environment among member states by January 1, 2010. Effective immediately, the agreement calls for all ASEAN members to give each other Most Favoured Nation status, effectively endowing preferential investment privileges upon all ASEAN states equally. The agreement also calls for ASEAN members to start gradually removing investment barriers to allow a much freer flow of capital and skilled labour by the 2010 deadline.
By that date, ASEAN has committed itself to treating all member citizens as national citizens for investment purposes. It has also agreed to open up all industries completely to ASEAN investment by the 2010 deadline, with a goal of opening up all industries to foreign investment by 2020.
The establishment of the AIA will be overseen by an AIA Council made up of ministers responsible for investment. ASEAN ministers signed the document into being at the 30th ASEAN Economic Ministers conference in Manila, but they refused to answer questions about the framework agreement until the closing news conference on Thursday. But throughout the meeting, ASEAN representatives have stressed the important role the AIA will play in opening up and stimulating investment in this region, which is now struggling with the aftershocks of the 1997 financial crisis.
ASEAN groups Brunei, Indonesia, Laos, Malaysia, Myanmar, the Philippines, Singapore, Thailand and Vietnam.

Based on this the system is able to establish the following conceptual relationships that together constitutes the context:

```
Investment: Investment purpose
Investment: Investment barrier
Investment: ASEAN investment area
Investment: Framework agreement - ASEAN
Investment: Minister conference, Asian minister
Asian group: Brunei, Philippines, Myanmar, Vietnam, Singapore, Thailand, Laos, Malaysia, and Indonesia.
AIA: Investment, region
```

MÍMÍR technology would pin down the following names and check their relationships with other contextual elements:

```
Manila
ASEAN
AIA
```

Manila is associated with signing. ASEAN is connected with such info as conference opening and nations. AIA is closely related to “members”.

2.2.7.5 Thematic interpretation

The MÍMÍR technology is also able to determine the governing theme of an article as the one shown in the previous paragraph. The news story from Manila would have been classified in terms of:

```
Framework agreement
ASEAN investment
```

Although always subjected to subjective judgement, it is not difficult to agree on the output given here.

2.2.7.6 Information distribution

Ontologies embedded in text will define its content. Central concepts constituting the core can be listed once the document is analysed. This summary function provides a simple, but important overview of the text found and can serve as a useful "super abstract". In the event that agents work on behalf of the user it is important that the analysis and match operation is made transparent. It is important that the agent conveys its findings and the match results in a way that enables the user to make decisions based on that. Explanations are central in that respect. The shared ontology between the analysed document and the interest model defines the rationale for why a document is important to read. The indexing system is non-binary. In other words it provides an analogue representation of the fit between interest and document. This has been exploited in order to show how the information is distributed across the document. A simple histogram will tell the user what part of the document is pertinent to the issues that he is interested in knowing more about (see figure 11).

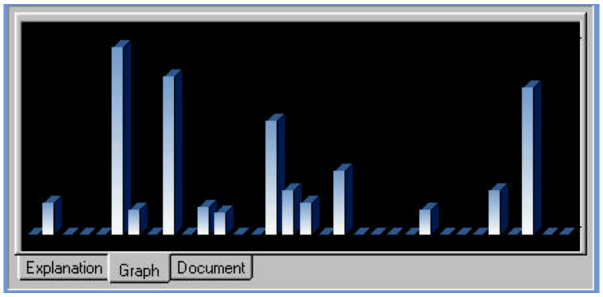


Figure 11: Histogram showing where the desired content in the document can be found and to what degree it is pertinent. Top of document is to the left. End of document is on the right hand side of the diagram. If a bar is selected and double-clicked the system will launch the paragraph represented by the bar.

2.2.7.7 Simple summarisation

At the time of writing the least developed part of MÍMÍR is the summarisation function. However, it is possible for the system to quote the paragraph that most closely matches the general interpretation model defined in terms of MÍMÍR. The Manila story would thus read in the news summary:

"MANILA, Oct 7 - The Association of South East Asian Nations (ASEAN) signed a framework agreement on Wednesday to create an ASEAN Investment Area (AIA) to stimulate investment in the economically battered region"

2.2.7.8 Applications

2.2.7.8.1 Introduction

CORPORUM may serve several purposes. It is typically useful for knowledge management purposes in order to use a corpus of documents as a knowledge base to support both problem solving and learning. Hence it should serve an important role within the Intranet. How this can done is described in detail in (Bremdal, 1999).. It should also be a candidate for E-business applications and serve various types of Internet services. Portal builders would certainly benefit from the use of CORPORUM. We will briefly address this. Also application of the various components may enhance the performance of classic document management systems as well as ERP applications.

2.2.7.9 Alleviating search on the internet

2.2.7.9.1 Front end to search engines

CORPORUM can in its simplest form be used as a front end to search engines. Given an interest model it is capable of feeding search engines with a set of keywords based on the thematic focus that resides in the interest model and post-process the results that these engines produce. Knowing that search engines like Alta Vista and Lycos are prone to how you specify a query this application can be very convenient. When constructing a query most search engines will be sensitive to the selection of both key words and in which sequence you enter them. For any given query the probability that a very good hit will show up among the first 10 is close to 1. However, the likelihood that all the most relevant can be found among the first 10 (or say 100) is very small. CORPORUM can be set up to process a theme by pushing various combinations of the same query. Several hits can be analyzed in parallel and ranked accordingly. This secures a very systematic approach and can be supported with no human intervention underway.

2.2.7.9.2 Portal building

Search is necessary when you do not have sufficient knowledge of what you are looking for or where interesting things may be located. A portal is a hub that brings links from several URL’s together. A portal may take considerable effort to build, but once it is there URL’s that belong to the same theme or have the same focus can be accessed by a simple click. Portals will provide a simple passageway to a specific domain or trade. A good portal can serve both news functions as well as e-commerce functions. However, a good portal needs to be maintained so that it does not point to dead links – URLs that no longer contain valid content. CORPORUM can be set up to create and to maintain a portal. Agents may nominate new and interesting links. These links can be included directly or be handled manually through a best-link nomination facility.

2.2.7.10 News and personalized information feeds

2.2.7.10.1 Customizing a news feed function

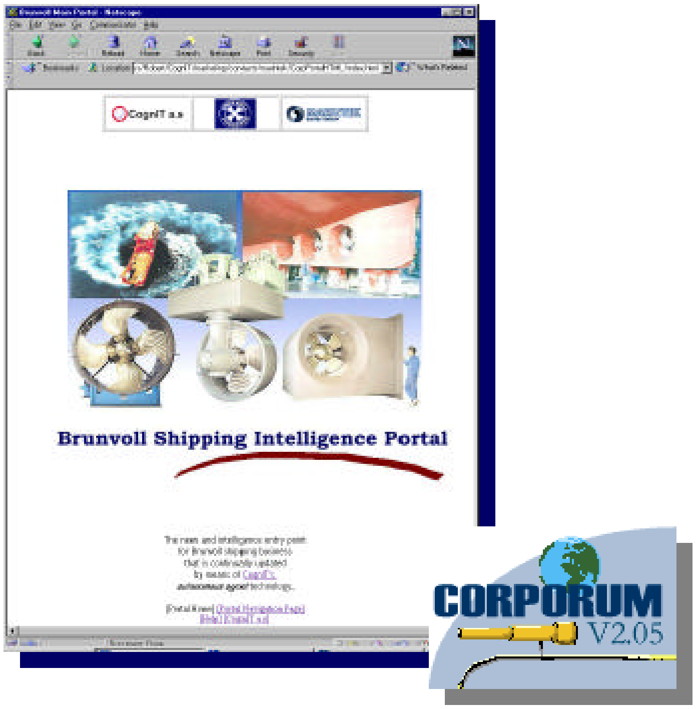
Figure 33: CORPORUM can be used to monitor news stories, conflict areas and political events.

By defining agents that focus on various subjects of interest it is possible to build a personalized newspaper functionality. Headline news can be fed in from several on-line media services from all over the world. Interest models could be set up for business, politics, sports, entertainment and similar. The user will be the chief-editor and the agents will be the journalists. Through agent training it is possible to customize focus so that events that are most important to you will be rated more important than others. An agent could be set up for sports while emphasizing news about the soccer team of your heart. An agent could monitor the Arab-Israeli conflict It would automatically fill in news about the involvement of Yassir Arafat, but restrain any accounts on his social life. However it would most likely volunteer news about analogue conflicts if you wish to receive this.

2.2.7.10.2 Public Information Systems

The age of Internet has generated a bias with respect to who will be updated and who will be not. Some groups in a population have access to the Internet and some do not. Some have the competence to surf out there many others do not. This division is strongly related to age. Older people have rare encounters with the Internet while younger groups apply this as a routine means of information access. Although this may be a temporary problem we have found that there exists a profound unbalance in terms of access as more and more governments and public services make public information available on the net. This can be alleviated through the use of CORPORUM. We have applied CORPORUM as a service for local governments. Information pertinent to the local community has over the years become available on the net. This information embraces news feeds provided by the community itself. It includes calls for meetings supervised by the municipality and it includes updates on rules and regulations. By defining the interest profiles of demographic groups CORPORUM has successfully demonstrated a news feed capability that is able to channel information about new information available on the net through E-mails and SMS messaging on the wireless telephone network.

2.2.7.10.3 Business intelligence


Brunvoll Main Portal - Netscape
CognIT a.s
Brunvoll Shipping Intelligence Portal
CORPORUM
V2.05

The news and intelligence entry point for Brunvoll shipping business that is continually updated by means of CognIT's *autonomous agent* technology.

[Portal Kiosk] [Portal Navigation Page] [Help] [CognIT a.s]

Figure 34: A business intelligence application for the shipping industry.

Summer 2000 CognIT has launched a particular version of CORPORUM called the *CORPORUM Business Intelligence Portal*. This is an application that supports expert agents that will systematically collect information about a certain market and market opportunities. It may solicit notes referring to two parties meeting and rumours about upcoming events such as the signing of an agreement or publication of a new and competing patent. The use of this type of business intelligence can be found in a separate description (cf. Bremdal, 2000). Figure 34 shows the front page for an application of the Business Intelligence Portal for company working in the shipping industry.

2.2.7.11 Virtues and shortcomings of CORPORUM technology.

CORPORUM as currently available, shows a rather high functionality. The components the system is made of are useful in many scenario's where Knowledge Management is asked for. However, where the CORPORUM components are able to capture thematic contexts, there are cases in which intentional knowledge is asked for. CORPORUM is not always capable of capturing such intentional knowledge. The main cause for this will presumably be the fact that no extensive world and discourse models are generated, on the contrary, CORPORUM aims at capturing contextual knowledge instead of performing deep semantic analysis according to (predefined) world models.

Currently there is much more knowledge and information available from analysis processes than the information that is communicated to the user. Obvious improvements of the system as it is now are visualisations of Central (or core) Concept Areas, allowing for browsing through document sets by using maps (cf. 2.2.4 on Cartia) and extending its linguistic capabilities to deeper understanding of smaller discourses (cf. 2.2.2 on AskJeeves).

3 Discussion

The current report discusses a large variety of approaches to natural language processing from diverse fields.

First the discussion descents into the history of the studies of natural (often referred to as "human") language, discussing different ideas, schools and initiatives.

On top of that there is a discussion on those stages in classical natural language research that are most basic to nearly all of the natural language processing systems as of today, as well as alternative approaches to natural language processing.

As means for the comparison of the different approaches, it is aimed to distill a model containing dimensions that will hold on nearly all approaches that were encountered. It is even more interesting to see that this model also shows an ordering of necessary-steps on which there seems high consensus in the research community. It is a model that is also used as a framework for reference by members of the international research and development communities that are not regarding themselves as being part of the "traditional" linguistic research paradigm.

This brings us to an interesting point in the discussion, the ways in which the approaches differ in their perception of discourse, thematic context analysis and intention. A rough separation can be made between the ways in which disciplines like linguistics, artificial intelligence, cognitive sciences and philosophy regard semantic and meaning, thereby heavily influencing the research conducted in that field. However, every of these research fields typically show a division in different NL paradigms as well, something which makes us believe that analysis of isolated research fields does not deliver valuable results.

Whereas in the field of linguistics there are discussions on the differences between (and needs for) "shallow" and "deep" processing of texts, at the same time there seem to be fields of research which originated from other disciplines (like f.e. statistics or connectionism) that made up their own mind, agreed upon the necessity for capturing deeper knowledge from texts, but focussed on a completely different type of representation mechanism for that.

The topic of representation mechanisms is a topic that is taken up by all researchers and applications in natural language processing, since it is the whole core idea of NLP; getting something more meaningful out of strings of characters. People with a pure linguistic background often see "meaning" or "semantics" as a formalisable entity that, on top of that, also has objective parameters. Principally they feel that approaches that can grasp the grammar of language should hold for description of semantics as well. The general consensus seems to be that not being able to capture at least a very general "universal" grammar in order to represent utterances means that there is no means for representing the "meaning" behind it either. This "Chomskian" tradition has as basic principle a simplistic but attractive reasoning: if people can theoretically utter a fast infinite amount of character and character-chunks (aka words), but instead of doing so focus on a small subset thereof, it can be concluded that there is some systematic order or "formalism" behind language. A typical proof for this is that people can even understand sentences that are never before encountered by a particular individual. Therefore something as a "universal grammar" should exist that makes up words and combinations thereof. This implies that it should be possible to create such a grammar.

Several attempts to the definition of such grammars are made, whereas "constituency grammars" and "functional dependency grammars" are the most commonly found grammar representations.

Besides of this paradigm in which one tries to exactly specify "what is going on" and which exact relations, types, transactions and movements are made in a particular world, there are other researcher who feel that these "universal representations" require too much predefined knowledge and too deep analyses of the world in order to be able to make them function properly. Such alternative approaches seek to specify semantics based on sub-symbolic information rather than symbolic representation mechanisms. Behind such approaches there is a fundamental difference in opinion on how a discourse should be represented. Should discourse be represented using explicit, human readable and interpretable formalisms (frames, logic, etc.), or is it that discourse is so context specific that it is only “learnable” from situations and history thereof and should it be represented sub-symbolically (like in connectionist models)?

In this learning paradigm there are approaches combining symbolic representation of elements with statistical learning of discourse (or context) from documents. Doing so has as disadvantage that no explicit representations of discourses are generated, while the representation can be much more efficient and faster to calculate, based only on properties that are described in a (set of) documents. More explicit knowledge (f.e. on relation types between concepts) can be extracted from such representations, but the basic approach is that the representations are not necessarily inspected by an external reviewer. I.e. the semantics of a specific narrative represented in such a way is “implicit”.

Now a major difference exists between several approaches to natural language processing on what it actually is that they represent. Whereas one approach clearly focuses on capturing the intention in a narrative, others concentrate on thematic context descriptions for such texts. As the former intent to capture both expressions on meaning of the narrative in the world, the intention that is expressed in it, and in which context it should be read and understood, the latter often presumes that the “reason of existence” of the text is already known.

Naturally, capturing all knowledge on why a narrative exists (who is it targeting, and why), what it says (what is actually topic, which actions or functions are described) and the context in which it is to be interpreted (does this text refer to woods in a European setting, or a South-American setting? Is this a park like an English garden, or a rain-forest that became a national park?) is a much more demanding task as the task of interpreting the context alone, given a certain presumption on the other two questions. Generally seen, representations in Information Extraction depend on the types of knowledge that have to be captured (is there a need for capturing intention, transactions or is there a need for capturing context), and the paradigm in which one thinks. AI people tend to think in frame-based or logic representations, statisticians in terms of information theory or probabilistic networks, and so on. Following such discussions can be a really worthwhile investment of time. An observation that can be made is that statistic and sub-symbolic approaches are often found in the “thematic context” description paradigm, where no back ground knowledge is used that represents knowledge about the state-or-the-world. AI approaches (knowledge based, logic based) are often directed towards the use of semantic back ground knowledge like for example the CYC world model (Lenat, 1996). In cognitive science the camps are really divided and whereas some tend to believe more in the sub-symbolic representations (statistics and connectionism are natural to these sciences), some others tend to believe more in making knowledge explicit, hence use representation formalisms that allow such interpretation. However, debates are fierce and one easily gets the feeling that they are not converging to a “single truth”.

This becomes even more obvious once regarding the applications and types of technologies used within such the products that are out on the market today. Often such products ally with one of the paradigms in natural language processing, and thereby are rather predictable in their strengths and weaknesses. However, sometimes an approach combines strengths of quite different approaches and thereby opens new ways of representing meaning and semantics of natural language documents.

When analysing products that can be found at the market nowadays, the first thing that pops the eye is the enormous role pragmatics play. Although research exists that performs rather deep and thorough research into the understanding of narratives in the broad sense, there is not much of this that is commercialized for use on what we regard as a very good test bed for natural language processing, namely the world wide web. Initiatives like CYC, which theoretically would help with exactly determining what the meaning of a specific narrative is, why it is and what the context is that it refers to, are simply to enormous to be used in “nearly” real time decision making machines.

Generally there are no approaches found that can deal with accuracy, understanding of both rational behind a document (a task which is performed reasonably well be humans), intention and context. Applications of natural language processing which aim at information extraction and retrieval range from simplistic approach as string- or pattern matching through shallow language processing to ever more demanding techniques. Depending on complexity and processing speed of computers, these approaches are useful in real time or are better applied off-line. Statistical and sub symbolic approaches seem to be more efficient in calculating and retrieving content as grammatical, deep understanding approaches are.

Which conclusions can be drawn from this? Can we draw lines to future development? Are there common elements in the different camps that should better be combined? Or should we proceed along the currently taken paths?

Of course, such questions are not easily answered. The diversity of the approaches undertaken really helped the field of linguistic research leap forward, due to fierce battling about rights and wrongs, pros and cons of the several approaches. On the other hand, there is also an argument that combining effort might help progress more. One of the reasons that this did not already happen is that the several disciplines have not reached consensus on which path to take. It is nearly certain that part of the discussion can be tracked back to the discussion in cognitive sciences on how to model human reasoning. The same camps can be found as in the field of natural language processing, and the same arguments are thrown. On top of that, pragmatic aspects of computability might have caused that "counting" or "statistic" and connectionist techniques became more applicable, thereby gaining efficiency by using less richer semantics. On the other hand, grammars are also perfectly calculable by Turing machines, but their richer representation makes them often unusable.

In the course of the current EU project "OntoKnowledge" the aim is to combine an explicit representation of the world to represent a specific part of a textually represented domain with techniques that are less domain dependent for extraction of the needed information. From the above one could easily conclude that this is not a regular approach to take. Nevertheless, it is expected that there are significant benefits to this approach of explicitly representing extracted information for user acceptance and understanding of certain parts of the world. Modern computing makes it possible to graphically represent much of the knowledge that is captured automatically today. It is possible to combine explicit, textual representation of (extracted) domain knowledge, graphical elements to navigate through knowledge and information and represent both a users interest, his information needs and answers thereon, with his or her place in a (virtual) world related to the knowledge that is present in it. This should be done in an intuitive, easily understood manner.

4 References

Aasen, I. (1848). Det norske Folkesprogs Grammatik. (In Norwegian) *For reference see Haugen, E.: The linguistic development of Ivar Aasen's New Norse* (2nd edition), Universitetsforlaget 1972.

Allen, J. (1994). Natural Language understanding. 2nd edition, Benjamin/Cunnings, Redwood City, CA.

Armstrong-Warwick, S. (1994) *Preface*. In Using Large Corpora (Ed. Armstrong). MIT Press.

Autonomy (1998). Autonomy: Technology White Paper. Autonomy Corporation, Cambridge, UK.

Besselaar, P. van den, & Leydesdorff, L. (1996). Mapping change in scientific specialties: a scientometric reconstruction of the development of artificial intelligence. *Journal of the American Society for Information Science.* JASIS 47(6) pp 415-436.

Bobrow, D.G. Fraser, B (1969). An Augmented State Transition Network Analysis Procedure. *First International Joint Conference on Artificial Intelligence*, Washington D.C. , 1969

Bobrow, R.J. and Webber, B.L. (1980). Knowledge representation for syntactic/semantic processing. *Proceedings of AAAI,* pp 316-323.

Bremdal, B.A. (1999) *Philosophy and its impact on our thinking about knowledge and language.* Technical Note. CognIT a.s.

Bremdal, B.A. (1999): "*Creating a Learning Organisation Through Content Based Document Management*", CognIT report #2.

Bremdal, B.A. (2000) MarCorp – Maritime CORPORUM. Sluttrapport for pilotprosjekt med Brunvoll a.s og MARINTEK, internal report (in norwegian), CognIT.

Briggs J. and David Peat, F. D. (1999). Seven Life Lessons Of Chaos. HarperCollins.

Brooks, R. A. (1991) *Intelligence without a representation.* Artificial Intelligence 47, Elsevier

Carr, H.A. (1931). The laws of association. *Psychological Review (38).*

Chomsky, N. (1957) *Syntactic Structures.* Mouton. The Hague

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* MIT Press

Chomsky, N. (1966). Cartesian Linguistics. Harper & Row Publ., New York.

Chomsky, N. (1968). Language and Mind. Harcourt Brace Jovanovich, Inc. New York.

Church, K.W. (1994). Mercer, R.L. *Introduction to the Special Issue on Computational Linguistics Using Large Corpora. In Using Large Corpora* (Ed. Armstrong). MIT Press

Croft, W.B. and Lewis, D.D. (1987). An Approach to Natural Language Processing for Document Retrieval. *In Proceedings of the 10th ACM-SIGIR Conference*, New Orleans, LA, June 3-5, pp 26-32.

DeJong, G. (1979). *A New Approach to Natural Language Processing.* Cognitive Science. Vol. 3, No. 3

Decker, S., Erdmann, M., Fensel, D. and Studer, R.(1999). Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information. *In R. Meersman et al. (eds.), Database Semantics, Semantic Issues in Multimedia Systems*, Kluwer Academic Publisher, Boston, 351-369,1999.

Declerck, T., Klein, J. and Neumann, G. (1998). Evaluation of the NLP Components of an Information Extraction System for German. *In Proceedings of the first international Conference on Language Resources and Evaluation (LREC),* Granada, 1998, 293-297.

Domeshek, E., Jones, E. and Ram, A. (1999). Capturing the Contents of Complex Narratives. *In: Ram, A and Moorman, K. (Eds.) Understanding Language Understanding: computational models of reading.* Bradford Book, Cambridge, MA.

Dreufus, H. (1979). What Computers Can't Do (Rev.edit.*) The Limits of Artificial Intelligence.* Harper Colophon Books

Duncker, (1935) Zur Psychologie des produktiven Denkens. Springer Verlag. Berlin

Eades, P. (1984). A hueristic for graph drawing. *Congressus Numerantium*, 42, pp. 149-160.

Fillmore, C.J. (1968). The case for case. In Bach E. & Harms R.T. (eds), *Universals in linguistic theory*. New York, Holt, Rinehart and Winston, 1-88.

Francis, W. and Kucera, H. (1982). *Frequency Analysis of English Usage*. Houghton Mufflin.

Goldfarb, L. and Deshpande, S. (1997). What is a symbolic measurement process? *Proc. 1997 IEEE Conf. Systems, Man, and Cybernetics*, Vol. 5, pp.4139-4145.

Gudwin, R. and Gomide, F. A Computational Semiotics Approach for Soft Computing. *Proc. of the SMC'97, IEEE International Conference on Systems, Man and Cybernetics*, Orlando, FL,USA, Oct. 1997, pp. 3981-3986.

Gudwin, R. (1999) Computational Semiotics: An Introduction. White Paper, University of Brazil.

Haegeman, Liliane (1991). Introduction to Government and Binding Theory, Basil Blackwell, Oxford.

Hahn, U. and Strube, M. (1996). ParseTalk about functional anaphora. *In McCalla, G. (Ed.): Advances in AI. Proceedings of the 11th Biennial Conference of the Canadian Society for Computational Studies of Intelligence (AI '96).* Toronto, Ontario, Canada, May 21-24, 1996. LNAI ,1081 , pp. 133-145. Springer Verlag. Berlin.

Halliday, M.A.K. (1970). Functional diversity in language as seen from a consideration of modality and mood in English. *Foundations of Language*, 6, 322-361.

Harris Z.S. (1961). Structural Linguistics. Chicago. University of Chicago Press.

Hirst, G. (1981). Discourse oriented anaphora resolution in natural language understanding: A review. In: AJCL, 7(2), pp 85-98.

Hirst, G. (1987). Semantic Interpretation against Ambiguity. New york: Cambridge University Press.

Hudson, R. (1990). English Word Grammar. Basil Blackwell, Oxford.

Inxight (1998). White Papers on Inxight product suite. Available from Inxight Corporation, Palo Alto, California.

Jacobs, P (1992) (Ed.) *Text Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, Lawrence Erlbaum, Hillsdale.

Jacobs, P. (1992). Introduction: Text Power and Intelligent Systems. *In Text Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval,* (Ed. Jacobs) Lawrence Erlbaum, Hillsdale.

James, W. (1890) Principles of Psychology. Holt, New York.

Kohonen, T. (1997). Self-organising Maps. Springer Verlag, Heidelberg.

Kowalski, R. (1979) *Logic for Problem Solving*. North Holland, New York.

Kristen, G. (1993). Kennis is macht. Academic Services.

Lamping, J., Rao, R. (1996). The Hyperbolic Browser: A Focus and Context Technique for Visualizing Large Hierarchies. *Journal of Visual Languages and Computing*, 7(1), 33- 35.

Lange, T.E., Wharton, C.M. (1999) Retrieval from Episodic Memory by Inferencing and Disambiguation. *In Understanding Language Understanding.* Ed(Ram, A. and Moorman). Bradford MIT.

Langston, M.C. Trabasso, T. and Magliano, J.P. (1999). *A Connectionist Model of Narrative Comprehension.* In Understanding Language Understanding (Ed: Ram and Morrman). Bradford MIT

Lashley (1960). The neuropsychology of Lashley; selected papers, edited by Frank A. Beach New York, McGraw-Hill.

Lebowitz, M. (1983). *Memory-Based Parsing.* Artificial Intelligence No. 21, Elsevier.

Lenat, D. B. (1995). CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, **38**(11), pp 33-38.

Luhn, H.P., (1958).'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, 2, pp 159-165.

Mallery, J.C.M. (1988). Thinking About Foreign Policy: Finding an Appropriate Role for Artificially Intelligent Computers. *The 1988 Annual Meeting of the International Studies Association*, St. Louis, Missouri.

Manning, C. D. and Schütze, H.(1999) Foundations of Statistical Natural Language Processing, MIT Press.

McCarthy (1954) Language development in children. *In L. Carmichael (Ed.), Manual of Child Psychology.* New York.Wiley.

Miller, G.A. (1956). The magic Number Seven, Plus or Minus Two. Some limits on our capacity for processing information. *Psychological Review*, no. 63.

Miller, G. (1995). WordNet: A lexical database for English. *Communications of the ACM* **38(11**), pp 39-41.

Minsky, M. (1966). Artificial Intelligence. *Scientific American* Vol. 215 No.3.

Minsky, M. (1975). A framework for representing knowledge. In The Psychology of Computer Visions (Ed. Winston). McGraw-Hill

Neijt, A. and Bakker, D. (1990). Computer Linguistiek: een overzicht in artikelen. Foris Publications, Dordrecht, NL.

Newell, A. and Simon, H.A. (1972). *Human Problem Solving*, Prentice Hall.

Newell, A. (1980). Duncker on Thinking. An Inquiry into Progress in Cognition*. Report CMU-CS-80-151– Computer Science Department*, CMU, Pittsburgh

Noble, H. (1988). Natural Language Processing. Blackwell Scientific Publications, Oxford.

Nonaka, I. and Takeuchi, H. (1995). The Knowledge-Creating Company. Oxford Reference Book Society.

Quillan, R. (1969). Semantic Memory. *In Semantic Information Processing* (Ed. Minsky). MIT Press, 1969

Ram, A. (1999). A Theory of Questions and Question Asking. *In Understanding Language Understanding Ed(Ram and Moorman).* Bradford MIT

Ram, A. and D. Leake, D. (1991). Evaluation of Explanatory Hypotheses. *Proc. of the 13th Annual Conference of the Cognitive Science Society*, Chicago, IL, August, pp. 867-871.

Ram, A. and Moorman, K. (1999) Introduction: Toward a Theory of Reading and Understanding. *Ed(Ram and Moorman).* Bradford MIT

Richardson, R. and Smeaton, A.F. (1995). Using WordNet in a knowledge-based approach to Information retrieval. *Technical Report CA-0395, School of Computer Applications,* Dublin City University, Dublin, Ireland.

Riesbeck, C.K. and Martin, C.E. (1986). Direct Memory Access Parsing. *In Experience, Memory, and Reasoning (Ed. Kolodner, Riesbeck).* Lawrence Erlbaum.

Rijsbergen, C.J. van (1979). Information Retrieval. second edition, Butterworths, London.

Rumelhart, D.E. (1975). „Notes on schema for stories." *In: Bobrow, D. and Collins, A. (eds.) Representation and Understanding.* New York.

Salton, G. (1968). Automatic information organization and retrieval. McGraw Hill, NewYork.

Shank, R.C. (1975). Conceptual Information Processing. North Holland, Amsterdam 1975.

Schank, R. and Abelson, R. (1977). Scripts, Plans, Goals and Understanding. Hillsdale, NJ: Lawrence Erlbaum.

Schank, R.C. (1982). *Dynamic Memory,* Cambridge University Press.

Shannon, C.E. (1948). The Mathematical Theory of Communication. The Bell Systems Technical Journal (27), pp.379-423.

Simon, H. A. (1976). *Administrative Behavior*. McMillan.

Skinner, B. F. (1957). Verbal behavior. New York: Appleton-Century- Crofts.

Sparck Jones, K. (1984) Automatic Search Term Variant Generation. Journal of Documentation, 40, pp 50-66.

Staab, S. (1999). Grading Knowledge: Extracting Degree Information from Texts. *Lecture Notes in Computer Science (1744)*. Springer Verlag, Berlin.

Studer, R. (19??). OntoBroker System.

Tegenbos, J. (1997). My Kingdom for an Agent? Evaluation of AUTONOMY, an Intelligent Search Agent for the Internet. In: Online & CDROM Review, 21(3), pp.139–148.

Thorndike, E.L. (1898). Animal Intelligence. *Psychological Monograph, no. 2, 1898. (Reprint School of Psychology C-MU 1986).*

Thorne, J.P., Bratley, P. and Dewar, H. (1968). The Syntactic Analysis by Machine. *In D. Michie (Ed.) Machine Intelligence 3*, Edinburgh University press, Edinburgh, Scotland.

Watson, J. (1913). *Psychology as the behaviourist views it*. Psychological Review, New York, no 20.

Wilensky, R. (1983). *Planning and Understanding.* Addison-Wesley

Wilks, Y.A. (1972). Grammar, Meaning and Machine Analysis of Language. Routledge and Paul Kegan, London.

Winograd, T. (1972). Understanding Natural Language. *Cognitive Psychology No.3*, Academic Press, New York

Winograd, T. and Flores, F. (1986). Understanding Computers and Cognition: A New Foundation for Design. Norwood, NJ: Ablex.

Winston, P.H. (1992). *Artificial Intelligence*. 3rd Ed. Addison-Wesley.

Wiseman, J. (1993). The SAS survival Handbook. HarperCollins Publishers, Glasgow, UK.

Wittgenstein, L. (1968). Wittgenstein's Note for Lectures on "Private Experience" and "Sense Data".

Wolff, J.G. (1991). *Towards a Theory of Cognition and Computing*. Ellis Horwood, Chichester.

Whorf, B.L., (1956). Language, thought, and readily. Boston: MIT Press.

Quillian, M. R. (1968). Semantic Memory. *In M. Minsky (ed.), Semantic Information Processing. MIT press*. Cambridge, MA.

Zarri. G.P. (1997) Natural Language Processing Associated with Expert Systems. *In Liebowitz (ed.), J: The Handbook of Applied Expert Systems*. CRC Press.

Zipf G.K. (1935). The Psycho-biology of Language. Boston. Houghton Mifflin.

Zipf. G.K. (1949) Human Behaviour and the Principle of Least Effort. Hafner. New York.

Appendix A: List of evaluated approaches and products

Overview of systems that were considered for this report. The reason that not all the systems are discussed in detail is not that they were not interesting, but rather due to the fact that this report aims at providing a state-of-the-art rather as an extensive system description. It was tried to cluster the systems according to their functionality and discuss one system from every of these groups.

Products and Software solutions:

- Ö **AskJeeves**: Full text questioning (one sentence)
- Ö **Allaire, ColdFusion (see Verity)**
- Ö **Autonomy**: KnowledgeServer
- Ö **Cartia**: Document Classification using an Information Landscape
- Ö **CoNEXor**: linguistic software (taggers, parsers etc.)
- Ö **Corporum/ BIP**
- Ö **DataWare**:
- Ö **DocNet**:
- Ö **Excalibur Technologies**: innovative Knowledge Retrieval software solutions (text, images, video), based on o.a. full-text search and Online Demo shows obvious WordNet usage.
- Ö **IBM Intelligent Miner for Text**:
- Ö **InfoMation**:
- Ö **Inxight**
- Ö **KISS**: Integrated approach to BP modelling and Document management (incl. deep text analysis)
- Ö **Knowlix**:
- Ö **LexTek**:
- Ö **Lingsoft**: Linguistic software, many nordic languages. Has norwegian thesaurus.
- Ö **MicroSoft:** MS' Digital Nervous System
- Ö **iKnowledge**: PerformanceWare
- Ö **NewsReal**
- Ö **OpenText.com**
- Ö **ORACLE/ConText**
- Ö **Orbital Products**: KnowledgeWare
- Ö **PCDOCS/FULCRUM**: knowledge management and information management (incl. FT-search)
- Ö **Quercus**:
- Ö **SEMIO**: Text analysis and search, automatic categorisation and indexing (incl. hypertext links)
- Ö **Software Innovation**
- Ö **Trion**:
- Ö **Verity**:
- Ö **X-Portal:**

Projects (mainly non-commercial or academic)

- **ADRENAL (Croft) (besteld)**
- **AutoSlog**
- Ö **CIIR** (Center for Intell. Info. Retrieval) university of Massachusetts
- **CRYSTAL (+Webfoot)**
- Ö **DECOMATE-II**: Virtual Library Project sponsored by EU
- **HASTEN/WHISK**
- **LIEP**
- **PALKA**
- **RAPIER**
- Ö **Sheffield University (NLP group):** Information Retrieval and Extraction
- **SRV**
- Ö **TreeBank approach:** Tagging for English Text Corpora. Supported by Dictionary.
- ▪ **WRAPPER** (induction) Systems (Kushmerick WIEN, 1999?)

Appendix B: Tests on Autonomy technology

This appendix discusses tests that where performed on Autonomy in the past (Tegenbos, 1997) and currently. The basic technology of Autonomy seems not to have changed too much inbetween, most of the observations that hold in 1997 still hold. The current tests where performed on portals utilising Autonomy’s web server solution, usually in the context of searching eBusiness data bases on relevant products for a customer of the portal. An example of such a portal is the Yattack portal (http://www.yattack.no), which was powered by the Autonomy engine.

Example:
Tested was an Autonomy implementation at www.yatack.com, an Autonomy powered Ecommerce portal. The gist of the test is that Autonomy slightly misses the point when they try to deliver language independent services. Autonomy seems to retreat to the point of doing real pattern matching, which actually causes an enormous loss of focus. The principle that Autonomy shows is that pure sentence parsing will not enable us to resolve ambiguities. Therefore systems like these need proper contextual knowledge. Autonomy obviously uses Bayes/Shannonian techniques for finding either statistically interesting concepts/words, or entropy based co-occurrences of words. When performing tests with Autonomy, one will find that a short question like: "I am looking for something that has to do with Motorcycles." will confuse the system and leads to results on hits on CDs, a MiniDdisc system and even a Calcium preparate. Unless the system has background knowledge about muscle-aches that have to do with riding an off-road for a while, there is no real relevance (maintaining such knowledge is not reported in the Autonomy case). Instead, Autonomy takes as input the phrase "looking for" and returns hits based on that.

Results on the question: “I am looking for something that has to do with Motorcycles”.

* AIWA AM-F75 minidisc med opptak
Helt ny modell fra Aiwa!
yaTack pris:2995.00

*CROSBY STILLS NASH & YOUNG Looking forward
Looking forward
yaTack pris:139.00

*DIANA KRALL When I look in your eyes
yaTack pris:139.00

*FOOD STATE Kalsium m/vit.D 200mg. 50 stk.
Nødvendig for elastisteten i muskler.
Viktig for skjelettet.
yaTack pris:169.00

The problem with accuracy seem to be inherent to probabilistic/information theory based systems and small data sets. As long as large volumes of text are present, it is possible to infer structure. But what if texts are rather small? Personally Shannon based decision tree learners perform rather adhoc/randomly on smaller data sets. Usually just accidental co-appearances of words are modelled, and this is something that can be observed in this example as well.