Charles University in Prague

Faculty of Arts

Master Thesis

2010 **Bc. Barbora Poláková**

Charles University in Prague

Faculty of Arts

Institute of Information Studies and Librarianship

Field of study: Information Science and Librarianship

Program of study: Information Studies and Librarianship

Master Thesis

Bc. Barbora Poláková

Natural Language in Web-based Search Engines

Přirozený jazyk ve webových vyhledávačích

Prague 2010

Thesis Supervisor: Mgr. Jan Brejcha

Thesis Oponent/Oponent diplomové práce:

Date of defense/Datum obhajoby:

Evaluation/Hodnocení:

Prohlášení (česky):

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

V Praze, dne 17. Srpna 2010

Statement (In English):

I declare that the thesis was prepared separately and exclusively with the use of quoted sources, literature and other expert sources.

Prague, 17 August 2010

……..……..……..

Thesis author signature /podpis diplomanta

Identifikační záznam (česky):

POLÁKOVÁ, Barbora. *Natural Language in Web-based Search Engines = Přirozený jazyk ve webových vyhledávačích*. Praha, 2010. s.79, Diplomová práce (Mgr.). Univerzita Karlova v Praze, Filozofická fakulta, Ústav informačních studií a knihovnictví 2010. Vedoucí diplomové práce Jan Brejcha.

Abstrakt (česky):

Diplomová práce je zaměřena na problematiku využití přirozeného jazyka ve webových vyhledávačích, zdůrazňuje multioborový charakter dané problematiky. Propojuje základní přístupy z lingvistiky, psychologie, kognitivní vědy, informační vědy a neuropsychologie, s cílem vytvořit její komplexnější obraz. Bližší pohled je věnován člověku jako hybateli celého vývoje v oblasti vyhledávačů a primárnímu uživateli takových systémů. "Lidský informační system" je prezentován jako vzor lepšího uživatelsky přívětivějšího systému pro zpracování informací. Charakter práce je plně teoretický a měl by sloužit jako základ pro další výzkum. Výsledkem je doporučení věnovat se vývoji plně vizuálních systémů s využitím znalostí lidského informačního system, který se na základě studia dostupných zdrojů jeví jako nejlepší cesta…….. [Autorský abstrakt].

Klíčová slova (česky):

Přirozený jazyk, umělý jazyk, zpracování přirozeného jazyka, kognitivní procesy, informační potřeba, lidský informační system, priming, vyhledávání informací, vyhledávací nástroje, vyhledávače, informace, znalost, reprezentace, webové prostředí

Descriptive entry(in English):

POLÁKOVÁ, Barbora. *Natural Language in Web-based Search Engines = Přirozený jazyk ve webových vyhledávačích*. Prague, 2010. pp. 79., Thesis (Mgr.). Charles University in Prague, Faculty of Arts, Institute of Information Studies and Librarianship 2010. Thesis Supervisor Jan Brejcha.

Abstract (in English):

This study is focused on the natural language exploitation in the search engines of the specific web environment. This paper points out the multideparmental character of the problematic and interlink the basic approaches from linguistic, psychology, cognitive science, information science and neuropsychology to create more complex image of the problematic. Closer attention is given to the human as the impetus of the development and as the primary user of such search engines. Human information system is presented as model for the better user friendly information processing system. The character of the paper is fully theoretical and should give the theoretical background for further research. As the result of the research is recommended focus on the fully visual approach in the development with the exploitation of knowledges about HIS.

…….. [Author‘s abstract].

Keywords (in English):

Natural language, artificial language, natural language processing (NLP), cognitive processes, information need, human information system, priming, information retrieval, search tools, search engines, information, knowledge, representation, web environment

Content

List of Figures

Preface

Natural language exploitation in information science is problematic which attend this department since the very beginning of its existence. Its multidepartmental character is the reason that makes it difficult to understand and research. And as in all multidepartmental studies there is lack of knowledge about the progress in other departments.

I had an opportunity to concentrate on the information studies from the socio-cognitive point of view during my studies in Finland at Åbo Akademi. The whole study program was significantly user centered and gave the opportunity to study the problematic from the other points of view than at the home university.

The systematic work on this paper thus started in Finland under the leading of Isto Huvila PhD (PD) and others. While continuing the research in Prague, I draw a lot from the Finnish experiences and information sources.

I found the user centered approach in studies very interesting and missing at our institute, and that's why I decided to write on such topic.

The English language is chosen because of the sources that are predominantly in English. However, the main reason was the possibility of the further studies in abroad for the PhD. The formal pages are written in both languages (Czech and English) and the text is in English only.

The references and bibliography is based on Czech standards ISO 690 a ISO 690-2.

I would like to thank the thesis supervisor Mgr. Jan Brejcha. Special thank belongs to Charles Nadeau for linguistic correction and PhDr. Markéta Opálková for formal corrections.

1 Introduction

Natural language as unique human phenomena always fascinated scientists in many different domains. This problematic has its roots in philosophy, psychology, linguistics, cognitive science and information technologies represented by artificial intelligence researches or information studies. Natural language clearly is a multidepartmental topic. That is, in each domain examined according to different approaches based on different domain related knowledge bases.

The natural language exploitation in information systems is broadly discussed in information studies and information technologies. However the results in these fields are still not satisfying. The reason is most likely related to the misunderstanding of the basic principles of human natural language processing and associated processes.

The attempt of this paper is thus to interlink the exploitation of natural language, thinking and information retrieval in use of web environment from the information studies point of view with the cognitive approach of the psychological domain by pointing out the main facts and assumptions. Subsequently, it strives to compare the consequential theories with the accessible natural language web search engines.

According to the evaluation of current natural language search engines, the aim will be to suggest the theoretical solution for a cognitively acceptable web based search engine.

The main goal of this work is to set often omitted basic characteristics of human language processing. The whole work is meant to be the theoretical background for further research on the human way of information retrieval, thought and its support by practical informational tools.

The complex theoretical background is presented in the six chapters. Each chapter corresponds to a specific domain and describes the problematic from its point of view. The final discussion underlines the multidisciplinary connection of the whole problematic and put it in context.

The first three chapters describe the information retrieval processes from the information studies point of view with description of basic search strategies. They are focused on basic principles description of information retrieval tools with focus

on web environment. The goal of these chapters is to provide basic theoretical overview of information retrieval tools functionality.

Chapter four describes the current natural language exploitation in web search engines. It contains the language definition and characteristics mostly from the linguistic and information studies point of view. The problematic of Natural Language Processing is also introduced. Finally, the end of the chapter also presents the typical example of natural language search engine called; Question answering search engine.

The largest part of this work is devoted to the human information system (see chapter 5) and cognitive processes (see chapter 6). This chapters attempt to describe important cognitive processes language related as well as the philosophical point of view represented by Wittgenstein´s representational theory.

1.1 General hypotheses

For an improved understanding of the content of this work, it is suggested to consider the following hypotheses that stand beyond the idea of development practical natural language web search engine.

1) **Do the natural language web search engines reflect the way of human's information acquisition? Do they reflect human way of thinking?**
    a. People usually do not think in whole sentences, but most likely in terms - keywords or images. To oblige people to think in whole sentences means to add new complex process in the whole system of information acquisition.
2) **Do natural language search engines serve as learning tool?**
    a. Natural language search engines give the right answer on right question. It does not enlarge the interest and does not give the alternatives or ideas for further work.
3) **Do human use language and text as the most common way for information processing?**
    a. People do not think in whole sentences, but most likely in keywords and images. The ideal approach would be the combination of most common ways of information processing: visual search + keywords (clustering) = mind maps.
4) **Is it valuable to cherish this natural language approach? Or is it better to focus on some other approach?**
    a. The natural language approach is important for the Natural Language Processing and its technical site, but it could never supply the schematized approaches of human information retrieval in the way of thinking.
    b. The natural language, visual and cluster search engines combination shows the best model of human's inner mind and ways of thinking.

2 Information retrieval process

"*Make the right information available to the right user, by analyzing content of information retrieval system (IRS) and user's queries to achieve the relevant of the items*"

(Chowdhury, 2003).

This work is based on hypothesis that Users should stand beyond the information retrieval techniques development. The reason is that information retrieval techniques are supposed to help users to orientate themselves across the reality. Their purpose is to find the relevant and pertinent information. In general terms; the purpose is to improve the way of living. Given this statement developers should find the way to fulfill this idea.

Information retrieval systems thus present the compact information environment consisting of all main components of IRS (Rosenfeld, 2000), where information retrieval processes are focused on searching and finding information contained in the contextual part of the information system environment.

The systematic **computer-centered** approach stands beyond the information retrieval system (IRS) strong technical support in the appearance of mathematical and logical algorithms for indexing and evaluation entities of the IRS (Yang, 2005) (see search engine decomposition). It is a significant challenge especially in the web environment (see chapter 3.1). This technical computer-centered approach was typical for 60s and 70s. Nevertheless in the 80s occurred higher interest in the cognitive aspects of information studies in general. This era is followed by the shift of interest from the technical background to the user's aspects. It is reflected in the cognitive approach of information studies. Later in 90s was this cognitive approach extended by the social aspects and user was set in the context of his reality, environment and other users.

User-centered approach is supported by many researches headed by Chowdhury (2000) or Rasmussen (2003) who presented the clear need to focus on users and adapt the IRS according to their needs and ability to use such systems.

As a proof that the user-centered approach leadership in development could serve the existence of the conception of Interactive web (Treddinick, 2006). And other related applications as for example systems as natural language IRS or visual IRS, that are supposed to be based directly on the users´ information needs.

2.1 Socio-cognitive approach

The user-centered IR research went through three main approaches: cognitive, social and socio-cognitive. Nowadays, the leading **socio-cognitive approach** represented mainly by the Danish and other northern schools on information studies (Hjorland, 2002; Ryutov, 2007). This theory estimates the importance of the user´s individuality as well as his social background and the contextual anchorage of user as well as the information.

This paper is specialized on cognitive aspects of this approach. The hypothesis considers the cognitive processes as the kernel processes for any IR process. In other words at the beginning of IR stands user respectively **users´ needs** and his ability to information acquisition, processing and behavior.

The aim of this paper is not to evaluate the complex socio-cognitive theory, but describe its cognitive part in consideration with its real exploitation represented by natural language search engines.

2.2 Information need

In general, the universal user requiring in IRS is considered as the need:

"*To find the right answer that would fill up the user's information need as fast and easy as possible.*"(Morville, 2006)

According to Case (2000), the information need is defined as recognition that user´s knowledge is inadequate to satisfy his aim. In other words could be user´s need defined as lack of information or knowledge considered during the appropriate cognitive processes.

Cognitive processes affecting the emergence of the information need are considered as results of **perception, communication or thinking process**. (Ingwersen, 1996).

Allen (2000) presents basic reasons that stand beyond the information need arise, that fully reflect the cognitive approach in the first two articles[1]:

1. *Failure of perception*
2. *Process of exploring a topic area so as the identification of alternative courses of action*
3. *Needs to associate alternatives with outcomes*

Since the cognitive processes in general are **dynamic and variable (**see chapter 6**)**, the information need has the same characteristic. The dynamicity and variability cause, that user seldom knows exactly what they need, because the potential information need is still changing.

However the changeability seems though to be partly limited. And the information need signifies the relative short term stability which enables User to focus on the solved situation and to achieve the goal during the concentrated proceedings. The concentration on the problem solving enables User to keep relatively stabile actual kernel information need. This relative stability allows the emergence of any IRS (Ingwersen, 1996).

Considering this, different approaches that reflect user´s effort to satisfy their more or less distinct information needs are rising. According to Morville and Rosenfeld (2002) there are two basic models of user information retrieval behavior: **browsing** and **searching**.

2.3 Browsing

Browsing, is a method through which users do not articulate exact queries, but find their way towards the desired goal (information) mostly accordingly by menus and links between interconnected system entities.

Primarily, the navigational system focused on problematic of user´s movement through the overflow of information representations in information retrieval system rather than the tool for information retrieval itself in the meaning of the finding and showing the result of search. Browsing is often connected to specific need **to learn** something about particular topic and not to search exact answer on question. This

[1] The last premise reflects the social part of the whole socio-cognitive approach.

need is based on the fact that the knowledge gap which creates the information need is wider. There exist different browsing strategies that suggest different information behavior of user (Morville, 2002).

Browsing activity stands on the border of **conscious and unconscious information retrieval** and it represents the overbearing information activity in the Web environment. This search strategy frequently presented phenomena described as "**topic drift**" effect (Yang, 2005). That is caused by changing the user's knowledge and experience basis during the browsing process, therefore changing the users´ further **motivation** and information **needs**. It is closely related to the cognitive processes of **knowledge representation (**see chapter 6.6**) and information reloading (**see chapter 6.5**).**

2.4 Searching

By contrast to browsing strategy, searching is characterized as exact setting of the query in the IRS, where the query reflects user's information need. Searching activity is fully conscious and the knowledge gap of information need is more specific and concrete. This enables users to create the exact query and get the exact answer. The query setting is based on specific chosen search strategy[2]. According to Chowdhury (2003) there exist three basic search types:

HIGH RECALL SEARCH - is defined by finding all relevant items in the stated topic. That means that the result offered by IRS will contain all relevant items reflecting the user's query.

HIGH PRECISIOUS SEARCH – according to this strategy, it is possible to find only relevant items, with a minimum number of non-relevant items. The result of the search does not have to be exhaustive, but the content has to be pertinent.

BRIEF SEARCH – contains just few relevant items and the major part of the result is irrelevant according to the presented query. In fact it is the opposite of the high recall search.

These three types of search strategy were established according to cognitive-user-centered approach to the user's information behavior. As it flows from the aforementioned definition of search strategies, this approach takes a holistic view

[2] There are different categorizations of search strategies; virtually each author has his own idea of this problematic and categorizations differed pursuant to the difference in the general approach. This paper considers the user oriented categorization of search strategies.

information retrieval problematic by taking into account not only the retrieval mechanism but also information needs, human-computer interaction during the search process and also the social and cognitive environment in which the process takes place (Chowdhury, 2003).

2.5 Seeking

Beside these two main and most often defined IR approaches, it is necessary to mention information seeking as the third kind of IR behavior. It is based on information needs of ordinary people (Case, 2006) or let say casual end-users (Kaufman, 2007).

Information seeking is, from all of the three approaches, the most unconscious. That means that users level of knowing what they want to find is the lowest of all. Users mainly promote the entertainment and sexuality related topics or other entertainment topics mostly expressed by short term interested topics (Spink, 2001). In general it is possible to include the seeking approach in browsing strategy as well as in search strategy. However, it is considered as the most incomprehensible and unmapped approach which reflects the majority of users' information retrieval practices in the web information environment. According to Spink (2001) has information seeking approach significant meaning for the further research.

Importance of information seeking is growing together with growing of the web search opportunities that reflect the user-centered approach.

3 Information retrieval tools

The main purpose of information retrieval tools is:

"*to make the right information available to the right user, by analyzing content of information retrieval system and user's queries to achieve the relevant of the items, regardless to the retrieval strategy*" (Chowdhury, 2003).

3.1 Web environment

The focus of this paper is on the World Wide Web based information retrieval system. Web environment has some typical characteristics that influence the IRS decomposition, above all: **size, dynamics, heterogeneity and accessibility** (Rasmussen, 2003; Yang, 2005; Morville, 2002).

3.1.1 Size

The size of web environment is enormous according to the traditional document collections (library, databases, and archives). On June 12th 2009, the size of web environment counted 119,974,427 active web pages, counted by the domains (Domain Counts, c2010). For instance the collection presented by Library of Congress contains approximately 120 million items and belongs to the biggest collection in the world(Library of Congress Home, c2010).

3.1.2 Heterogeneity

The important characteristic of the web collection is the heterogeneity of its items. The format variability of presented documents in web environment is wide and still extending[3]. The correlated factor that extends the size of web is the large redundancy of information which is caused by the heterogeneity of formats (Clarke, 2001). Redundancy is usually understood to be negative effect of huge document collection; nevertheless for natural language search engines, redundancy is one of the important indicators of answer's relevance.

[3] Let see the format characteristic of Dubline Core Metadata Element Set (DCMI Home, c2010)

3.1.3 Dynamics

Rasmussen (2003) presented interesting results of research carried out in 1991, where 99% of the observed web pages had changed after a year. In comparison to the traditional collections, that were relatively static, it is an underlying characteristic of web document collection.

3.1.4 Accessibility

The dynamics is strongly related to the accessibility in the connection to the problematic of web publishing. In general, the accessibility is the basic characteristic influencing the web environment development. Interactivity in format of free publishing is mainly perceived as a benefit, however on the other hand it causes problems with information organization and retrieval in such large, heterogeneous and dynamic collection. The way to deal with this web publishing noise could be found by following the recommendations of information architects, or content managers, and select exact responsibility for publishers (Brys, 2004; Batley, 2006). That could significantly help institutions and other directed organizations to constitute their web environment. Unfortunately, a huge part of the web environment is taken by single individuals. In that case, it is harder to apply content management.

These individual users generally have no idea about basic regularity necessary for web environment coherency, because of the voluntary character of these regulations. And thus information noise rise up as the result of user´s nescience and makes difficult systematical orientation in the web environment.

The research on the responsibility apprehension of web publishers is closely related to the problematic of users´ behavior in the web environment; however this topic is beyond the scope of this paper and shall be discussed elsewhere.

3.1.5 Scope

Another limitation of this paper is related to the domain orientation. The case of IRSs related to specific domain, where the research is limited by the knowledge base of the domain and users themselves. **Domain specific** systems (Thompson, 2005) could master the terminology and ontology which could bring successful results. Thanks mastered sublanguage specifics of the domain and it environment. This kind of system is than identified to the close specific group of users who are familiar with

the domain and its language system. These limits significantly ease creation of proper complex environment.

On the other hand **Open domain** systems provide more possibilities to compare, and thus for instance, facilitate the treatment of the traditional match approaches to find the right answer. Open domain involves diverse users base, because of its openness. Its goal is to catch all sublanguages that could appear in the affected environment and bring it in the context (Proudfoot, 2009).

In any case, it brings the research back to the effort of universe description. That is basically the goal of web environment especially semantic web (Treddinick, 2006).

3.2 Search tools

As it has been pointed out, users respectively their needs stand at the beginning of the information retrieval processing. The universal user's need in IRS is considered as the solicitude:

"*… to find the right answers that would fill up the user's information need as fast and easy as possible.*"(Morville, 2006)

The search tools can be divided according to the information retrieval strategies into navigation IRS and search engines.

3.2.1 Navigation

Navigation is the basic tool for browsing search strategy. It could appear in the form of directories, menus, links, FAQ etc. Their purpose is to offer the view on the ontology of hidden classification exploited in the system for indexing and information retrieval. Navigation system serves as the perfect tool for information retrieval without exact need which embodies browsing as well as seeking strategies.

Navigation tools represent important aspect of user's behavior in interactive web environment, where users are able to take part on the creating and editing the appearance of navigation (classification and indexing) according to their actual needs (Tredinnick, 2006).

Nowadays, it is usual to find methods of navigation in traditional search engines. These two information retrieval approaches are thus interconnected. Nevertheless,

the navigational approach, which seems to be closer to cognitive processes, is considered only as an auxiliary method.

3.2.2 Search engine

Search engine as the type of IRS is probably the most common information retrieval tool. Search engine as the query based IRS representing the searching strategies. However it is also significantly exploited for other information retrieval strategies as for instance browsing.

3.3 Search engine decomposition

Nowadays there is a myriad of different search engines in different information environments diverging in the size of the indexed collection, scope of the collection, interface, and mainly in technological support. However, these search engines are different, even in the same web environment, where they all perform the following basic structure (Chowdhury, 2001) (Figure 2).

3.3.1 Information environment point of view

Information retrieval system presents the complex information environment[4] consisting of basic entities: **user, system** and **content** domain (Toms, 2002) (Figure 1).

3.3.1.1 User

As noted by Toms E.G. (2000): "*Users bring to the process their human information processing capabilities*". Users are the reason of existing and developing information systems, therefore they are the dominant part of the search engine information environment.

The results of the shift from computer to user-centered approach can be seen in the existence of the "**interactive web**" or in the facilities and features of classic information environment – comments, profile options.

---

[4] For term "information environment" is often used equivalent "information ecology" used by Rosenfeld in Information architecture for World Wide Web (Morville, 2006).

The reasons why users are so important for the information environment seem obvious. Users stand literally at the beginning and at the end of all information activities of all systems. They activate information environment by information needs formulated in information tasks. They also stand at the end of the process, expecting appropriate output from the system, usable for them in the meaning of the information need complement and containing relevant valuable information. It has to be presented in format readable and understandable for users.

3.3.1.2 System

Systematic entity could be defined as "*a set of dynamic computer processes, contributing its artificial information capabilities*" (Toms, 2002). System element contains all technical and technological support as well as information about resources and all processes related to information as a product of the IE.

Toms (2002) defined this element solely on a technological basis while Rosenfeld (2002) considers it more extensively. In his point of view, this part of IE is rather defined as **context** and the subject of this element falls also outside of the technological area in the sphere of business organizational context. This contains all missions, goals, strategies, staff, procedures, budgets and culture. Morville and Rosenfeld (2006) submits the idea of dividing the traditional system concept in two rather independent parts: **technology**, which would reflect concept of Toms system element and **context,** which consist of business sphere and politics of the IE. Nevertheless, Rosenfeld has not developed this idea any further enter into details of this idea.

3.3.1.3 Content

The last element of IE embodies the information as the product. Content element consists of factual information (documents, applications, services and metadata) that user need to use or find (Rosenfeld, 2000). The granularity of the content information could differ according to the type of IE (document, sentences, phrases, words). They are organized, structured and contained within the "logical superstructure" of IE (Toms, 2002).

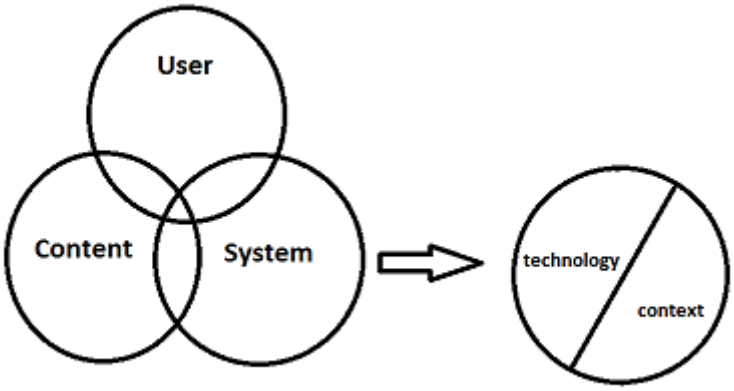

Figure 1 Information retrieval system entities

3.3.2 Systematical point of view

Search engines work on principle of automatically pre search. They automatically search the web environment, or some of its part, based on set of technical criteria. For this task search engines employ special program, called **spider**[5]. Spider is crawling over the websites and collect information found in the web documents in form of metadata (title, subtitles, metatags, author's keywords, full text and other position of relative importance on selected web pages). The difference between these programs is in the algorithm that established their motion in the web environment. When they find the sought-after information, they send the representation of the web page into the **index**, where are these metadata information stored and regularly updated.

From the side of the user, it is then possible to find the website that contains relevant information by using the word/phrases and their combination in the search engine interface. These words should fit into the indexed terms that allow the system to find the distinct connection to the websites containing these words or their combinations. The combination is supported by use of keywords and descriptors corresponding to the artificial language used for indexing and elements that determine the specific relation between these terms and limitations of the query. The most common is Boolean logic, but the possibilities are greatly wider (Chowdhury,

[5] Synonymous for spider are for instance bot, robot, trawler or indexer (Sklenák, c2009a)

2003). Finally the search engine offers the list of available websites, which is already arranged by an algorithm according to its relevance to the user's query[6].

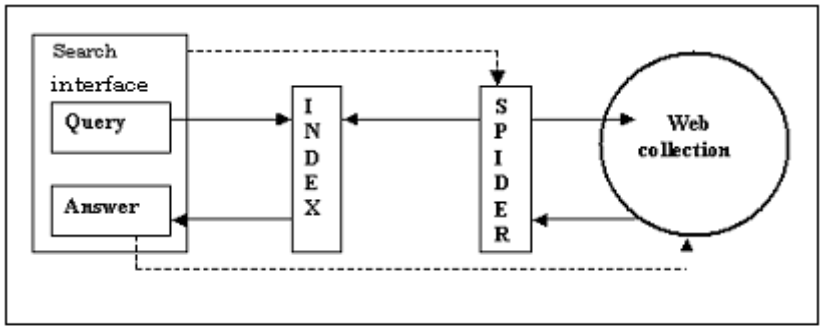

Figure 2 Search engine schema

The domain of web search engines and IRS in general is tremendous. Recently, the development has been significantly focused on an improved presentation of results and the method to measure the relevance of web items (Yang, 2005). However, the uprising of interactive web and semantic web theories increases the importance of the focus on user entity.

[6] These algorithms reflect the authority of the website, frequency of use and popularity, or the connectivity in the web environment as for instance *Hyperlink Induced Topic Search* (HITS) algorithm or Google´s *PageRank* algorithm (Yang, 2005).

4 Natural language in Web search engines

The following chapter is focused on the problematic of input and output of search engines with utilization of the natural language as the foremost language for user's query. Despite not being the main focus of the information studies, this problematic present a significant background of research, which roots gained upon the 60´s to the grounds of the artificial intelligence and culminated around year 2000. That time was presented MULDER as one of the first automated question answering systems available on the web (Kwok, 2000) and the development of question answering systems and application of **Natural Language Processing** (NLP) techniques was in the main focus of information studies presented by TREC experiments of the Question-Answering Track.

4.1 Language characteristics

The definition of language basic characteristics is a complicated task. Linguists, philosophers or cognitive psychologists usually have different viewpoint on the importance of language attributes. The following fundamentals characteristics are based on the studies of Brown, R. (1965), H.H.Clark and Clark (1977) or Glucksberg and Dansk (1975) as they are presented in Sternberg´s Cognitive psychology (2009).

1. Language is primarily intended for **communication** in the group of people who share the same language. Nowadays, this definition is extensible to the human computer communication, which is based primarily on the language.
2. Language is **arbitrary symbolical**; it adds the arbitrary relations to the signs, symbols and their meaning. The exact combination of signs or sounds (all words are symbols in any representation) has meaning following the conventional custom and creates the **substitute** expression to reality and abstract forms.
3. Language is **structured**. Modification applied on any structural level causes alternative comprehension to the language. Language structure has its own multilevel character consisting of syntax, semantics and pragmatics on the **vertical level** (Havlíčková, 2008).

- **SYNTAX** – Syntax is represented as grammar. Grammatical schema influences and forces relations between the vocabularies and, in connection with the context, they construct the framework for language.

- **SEMANTICS** –Semantics describes the meaning of the words, signs and other linguistics formations (it is related to the arbitrary symbolical characteristic of language)

- **PRAGMATICS** – Pragmatics represents the meaning of the words signs and other linguistics formations in their context.

Contrary to the schema of artificial language, its vertical structured schema applied on natural language is inadequately analyzable and definable as a result of its dynamics and contextual inconsistency.

On the **horizontal level** is language analyzed as sign, word, phrase or sentence.

4. Language is **generative** and **productive**; the derivation and production of new words, phrases and sentences is almost limitless.
5. Language is **dynamics**; each language is changing during the time and it does not matter what kind of language it is. The difference is in the tempo.

The aforementioned characteristics are valid for all kinds of language. The following part will be focused on the difference between natural and artificial language.

More information on language in connection to cognitive processes and Human information system is presented in the chapter 6.4.

4.2 Natural language

Natural language is traditionally defined as set of signs which is use by people to communication and is constantly developed by using. Language is based on informal and general accepted rules (Balíková, 2009).

4.2.1 Asymmetry

The basic characteristic of natural language is **asymmetry**. Asymmetry is based on synonymy, homonymy and other features, that cause some of the most significant problems of natural language usability in IRS. The problem of asymmetry is reflected mostly in the understanding of the context. Contextual relations may influence and change the whole meaning of the utterance (Dey, 2001).

4.2.2 Context

The contextual information contained in text and affecting the understanding is a consequence of the language structure application. In this case, it is articulate with the pragmatic part of language, which designates the meaning assigned to the language expression. This part of language has to be learned by users, to master their use of language. Another important attribute of pragmatic side of language is its dynamics. It is developed and changed during the time according to the environment.

4.2.3 Knowledge

Pragmatic aspect of natural language reflects the idea of knowledge, as defined in the theory of **knowledge society**. The main difference between information and knowledge is based upon its context.

Information in general is a contextual independent unit which can be indexed and organized according to norms and standards. Information is fully independent from its producer.

On the other hand, **knowledge** is based on contextual engaging according to its faculty to be defined as „*information in use*“ which is involved by experiences of author and specific development environment.

Knowledge takes place in human minds and is not necessarily expressed. In human mind is stored and organized in the form of knowledge structure (see chapter 6.6) which helps to understand and manage the interaction with reality.

Hypothetically the knowledge is in fact the pragmatic reflection of information presented by intellectual capital of individuals (Bukh, 2001). Knowledge is thus the basic product of human information system, rather than information which is in HIS automatically linked through the context.

4.3 Artificial language

Comparing all these general language characteristics with the artificial language, which is usually used as input in different search engines, it can be associated to the characteristic differences that cause the use of artificial language as query language and tool for searching.

First of all, artificial language is **symmetric**. Therefore, all items have defined specific meaning and relation that are stable, unequivocal and build upon the compact logical system of language. This structure also solve the problematic of context and pragmatic side of language which is, in artificial language, defined by relations and attributes representing the functionality of single sign (descriptors, keywords).

A fundamental characteristic of artificial language as supposed to natural language is its lack of dynamism. Changes are slow and come rarely. That makes the use of artificial language much stable then the use of natural language for all kinds of automatic processes.

4.4 Natural language processing

Domain of **Natural Language Processing** (**NLP**) explores the possibilities of natural language exploitation in information systems. The main focus is on the automatic analyzes of natural language texts for purpose of IR. These researches stand beyond the results in the field of automatic translation or, for the purpose of this paper, important query rewriting in natural language search engines (Hedlung, 2003). Nowadays, as a consequence of the acceptation of the interactive and semantic web concept, the idea of exploitation of natural language in search engines

is significantly reemerging. One of the results is the development of Natural language search engines represented for example by Wolfram-Alpha.

4.4.1 Natural language systems

Simultaneously to the overall effort to simplify the access and manipulation with the web content, arose attempts to apply the natural language as **query language** presented in user friendly interfaces.

Natural language as query language has different forms of presentation. Basically there are three forms of queries: **words, fragments, sentences**.

- **WORDS** established as query in search engine are already acceptable as queries in a majority of recent search engines that facilitate free text retrieval. They are based on perfect match, and thus, search for the same string of signs represented in the database. This is the easiest way to make use of natural language. Nevertheless the results are on low level of search recall. And it is not actually possible to talk about real language, because it does not engage all parts of language system. This means of query language only substitute keywords employed by ordinary query language – for example: „weather“
- **FRAGMENTS** are combinations of words with at least minimal exploitation of logical principles of natural language. They are represented by strings of words – for example: „weather in Canada”. As is obvious on the example, natural language fragments already contain kind of relation (logical and contextual) among words and represent the higher level of information expression. It actually anticipates the traditional process of information retrieval, where relationship between single keywords is based on exploitation of Boolean or other logical system. Fragments by contrast to words can be used for free text retrieval only without allowing keyword substitution. Their exploitation as keywords would need the destructuralization to the basic terms.

- **SENTENCES** are the most representative example of natural language queries. They are easy to identify because of the syntactic language part represented by punctuation and standardized rules, for instance, the first capital letter of the sentence. Problematic of sentences in use of search engines was developed and exploited in the special **Question-Answering systems** that facilitate their translation in the artificial query language (see later). An example of natural language search engine could be True Knowledge (True Knowledge, c2010) or lately very popular Wolphram Alpha (Wolfram|Alpha, c2010).

Despite being under active development and their deficiencies, these systems have a great potential to achieve the purpose of natural language search engines. It stands on a knowledge base created by users and the communication with the system is set in conversational mode, which presents a plausible progressive approach, predicting computers to be able to understand natural language and human reasoning by 2020 (Goh et al., 2007). Same author presents the web-based conversational system AINI, which is based on typed natural language conversation. In fact, it is not the full-value conversation, because of the lack of learning aspect which is a natural part of human communication. The divergences are solved by human interaction. Nevertheless, AINI´s results are impressive. This system has the ability to keep conversation on appropriate topic even though the human participant divaricates from the topic of the conversation. AINI thus keeps up the dialogical conversation between the user and the system, which could and should represent the next level of NLP IRS development.

The reason, why is AINI system so successful is the domain specification of its database. The goal of the semantic web and the web environment in general is to develop search engine which would be able to search enormous amount of data in the web environment. One approach is concentrated on domain specific, limited database and the second is based on open domain and enormous size of database. Both approaches have its explanation advantages and disadvantages.

4.5 Natural language search engine decomposition

Natural language search engines are efficiently represented by **Question-answering** (QA) systems. In general, the aim of the QA systems is to allow users to formulate their information needs in natural language and express it as a "normal" question in the interface of the information retrieval system. These systems differ in technical support like algorithm as well as in the way to present the results of search, character of the knowledge base and the whole system architecture.

For the QA systems, the web environment is a more effective option than the similar but relatively small and well organized document collection. The first reason is that the large information source will more likely embody the exact answer on the question in natural language. Therefore, it may more likely find the *perfect-match* search (see chapter 4.5.2). However, in the situation where there are no perfect matches according to the question, the redundancy and language variability, that are essential in the web environment, would facilitate the recognition of right answer by probabilistic methods, as the truth has tendency to appear more often than the nonsensical information in the content collection (Azari, 2004).

In general we could agree on three levels system decomposition, which are based on **query formulation, search engine** and **answer extraction**. Each subsystem will be described in following paragraphs.

4.5.1 Factoids

As was recently presented, the query transformation is exhaustive, but not impossible. QA systems are focused on factoids, by which focus on questions whose answer is clear and distinct. Thus these systems are engaged as the encyclopedic and cosmographic way of information provider.

At the moment when the QA system has to process the question of general and/or economic relevance, such as "*Can I travel with American pass in Mexico*?" most of the current QA systems are not able to answer it (Soricut, 2006).

4.5.1.1 Frequent Ask Questions

The solution proposed by Radu Soricut (2006) is the change of the knowledge base of such systems from the overall web environment to the document collection of

Frequent Answered Questions (FAQ). The original data collection is thus significantly narrowed; however it still corresponds to the web open domain environment.

The basic idea is, that FAQ contain most frequent questions expressed in pure natural language mainly the general, economic or other character that are not easily answerable in any accessible QA system via its logical system. The questions and answers in such collection are distinguished in particular **QA-pairs**. These QA-pairs are presented in different syntactically and semantically language formats. This characteristic presumes a high level of redundancy at this time considered as an advantage. Redundancy namely brings the variations of language utterances and specific forms of fact expression (Dumais, 2002).

It applies **two-step approach**, which is in first step oriented on recall. The aim is to retrieve most of the QA-pairs regardless of other irrelevant data. The second step is oriented on precision, while the previous result is filtered by using several logical limitations to reduce the level of noise.

As a result of the human interaction in FAQ base, it provides the current condition of natural language, because the whole base is created manually. Soricut (2006) suggests that the direction of research in the NLP should concentrate on exploitation of such actual knowledge bases as represented by FAQ. This suggestion reflects fully the user-centered approach of research. The research gets closer to the user, by analyzing direct "*normal*" language and not the published texts that could be influenced by stylistic demands of presented domain. That is important to distinguish, especially in the web environment, where the solicitude is to span all possibilities of language expression[7].

4.5.2 Query Formulation

Query formulation is one of the most challenging tasks of natural language search engines. The aim of the query formulation is to rewrite the question presented in natural language into the right query which is fully understandable for the search

[7] The other problem is the multilanguage characteristics of the web environment, however this topic is beyond the scope if this paper which is focused on the English as the most widespread language of web environment (Berendt, 2009).

engine and that will generate reasonable and relevant answers. That means reformulate the question in artificial language appropriate to the integrated search engine.

There are different ways to solve this task. For instance the AskMSR use parallel eight rewrite heuristics which deals with the problem of natural language asymmetry in different ways.

4.5.2.1 ANDing

The simplest method of query reformulation is *ANDing*. This technique rewrites the question in the string by using only AND between all words from the original question. For instance the rewritten question: "Who will be the next president?" would be reformulated in the following logic string:

who ***AND*** will ***AND*** be ***AND*** the ***AND*** next ***AND*** president

It is obvious that this method is limited by the Boolean logic itself (Peregrin, 2004) and the necessity to add another limitation as the *STOP word list*[8] to achieve a realistic answer.

4.5.2.2 Back-off strategy

An additional method used by AskMSR is the *back-off strategy*, consisting in the seeking for the exact phrases contained in the document collection. For example question: "Who killed Abraham Lincoln? "would be in this case rewritten in:

- <LEFT> "killed Abraham Lincoln"
- "Abraham Lincoln was killed by" <RIGHT>

<LEFT> and <RIGHT> determinate the side where is expecting the right answer (Azari, 2004).

[8] STOP word list is the list of signs that automatically omitted from the retrieval process, because They have no informative value (Sklenák, c2009).

4.5.2.3 Question Parsering

Question parsering determines the syntactic structure of the question. The parsing methods employ statistical techniques and use learned probabilities of word relationships to guide the search to the best parse. MULDER as the representative of the first trial of web based QA use for instance the **Maximum Entropy-Inspired** (MEI) parser (Charniak, 2000). MEI is based on probabilistic generative model; which represents the *top-down* parsering process, where the pre-assumption is, that the sentence always has a logical structure and is bearer of the semantic meaning.

The analysis starts on the top level embodied by sentence and continues down the lower lexical forms. This top down analysis finds its advantage in the simple detection of the constituents. Sentences are normally tokenized into words by the space between them. **Figure 3** presents the parser tree of the question: "Who killed Abraham Lincoln?" according to the Charniaks model[9]. The MEI method achieved 90,1% average precision/recall for sentences of length less or equal to 40 constituents in the test held on the standard close text test base (Charniak, 2000).

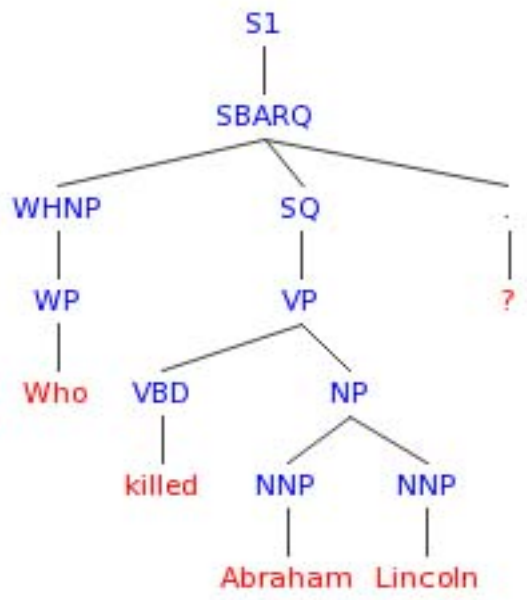

Figure 3 Charniak´s Parser tree

[9] The model was created by phpSyntaxTree v1.10. This software was developed in the National centre for language technology in Dublin, Ireland. Its demo is freely available on the web adresse http://lfg-demo.computing.dcu.ie/lfgparser.html .

Nevertheless, the parser is not able to recognize at first glance the right structure of the sentence. To resolve this situation, **multiple parse trees** are constructed simultaneously, capturing multiple hypotheses for an input string, based on a consideration of the likely different meaning that words in a phrase can have (Azari, 2004). The final choice of the parser tree is based on probabilities of each parser tree as the sum of probabilities of all nodes in the tree.

As the additive facilitation is for example in the MULDER´s parser integrated the lexical analyzer called **PC-KIMMO** which analyses words and tags previously unseen items that appeared in the user's question (Kwok, 2000). This process is set on the morphological tokenization on the word level (**Figure 4**). The question is therefore analyzed by means of the deepest morphological level and reflects the rules, lexicon and word grammar of English language (Antworth, 1995).

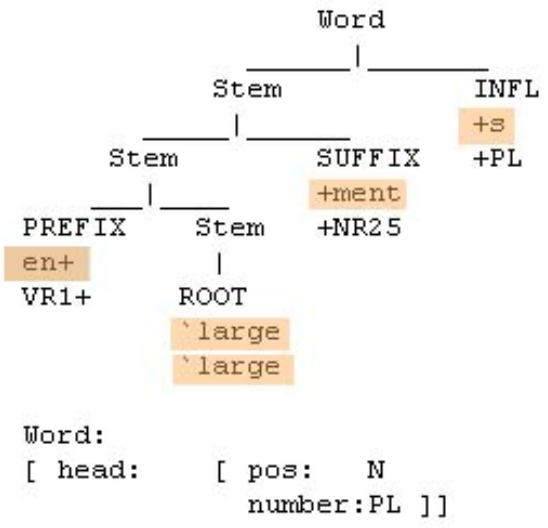

Figure 4 PC-KIMMO parser tree of word "enlargement "

4.5.2.4 Question classifier

After the making of the parsing tree, the whole question structure is analyzed by **Question classifier**. This part of search engine works as detector of three basic kinds of question:

1. NOMINAL – answers in form of noun phrases.
2. TEMPORAL – answers in form of dates.
3. NUMERICAL – answers in form of numbers.

Questions are distinguished mainly according to ***wh-phrases***. Wh- phrases quote the questions whose answers are other than Yes or No. Wh-phrases may be characterized as question operators "*What*", "*Where*", "*Who*" that represent the nominal questions. "*When*" is also accepted as wh-phrase introducing the temporal questions. Eventually the wh-phrase could be the entire phrase as for instance "*Which book have you bought?*" (Lai-Shen Cheng, 1997).

The numerical answers are easily determined by "*How much"* or "*How many"* phrases and the contained units of measure.

However some of question phrases quote ambiguous meaning of the sentence, which causes confusion of the system. In such situations system consults the question and its variants with **WordNet**.

4.5.2.5 WordNet

WordNet (WordNet, c2010) is exploited by various systems dealing with natural language processing. It is machine readable database developed at the Princeton University. WordNet represents semantic network of English language in the web environment. It contains words grouped into sets called *synsets*.

Synsets are linked to each other by different relations based on natural language asymmetry as for instance synonyms or hypernyms (Gentile, 2008).

Hypernym represents kind of semantic relation between terms based on semantic superiority (hypernym) and subordination (hyponym). For instance, Niagara Falls is a hyponym for the concept of *waterfall*, its hypernym (Snow, 2005). By using WordNet, question classifier is therefore able to determine the right meaning of the ambiguous questions.

4.5.2.6 MULDER query formulation

Query formulation in MULDER (Figure 5) employs more than one heuristic and thus is able to send to the search engine parallel seven different queries based on various principle of rewriting. The largest challenge in this part of transformation of the query is the extract of stop-words and reformulation the auxiliary verbs expression. Then the query is extended by alternative versions of words to cover the language asymmetry. Atomic noun phrases are clearly defined as inseparable. After

this exhaustive analysis is the question is transformed in the query with respect to its natural structure (Kwok, 2000).

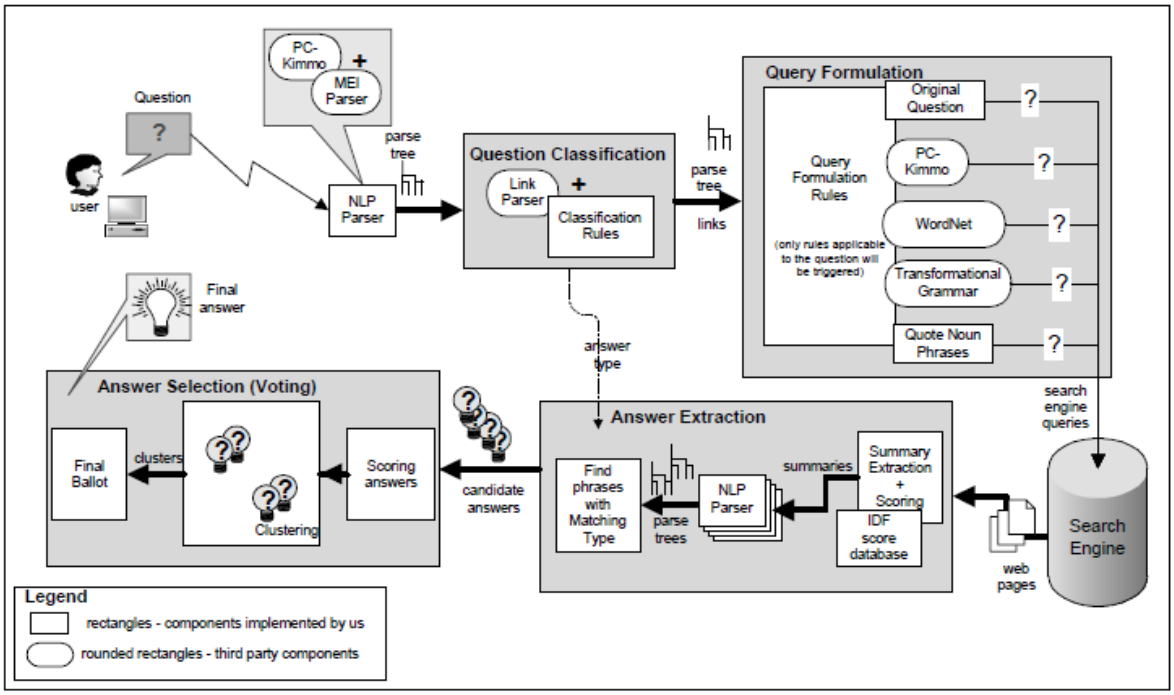

Figure 5 Architecture of Mulder according to Kwok (2000)

4.5.3 Semantic tractablility theory

PRECISE, another QA system, use for question rewriting **Semantic Tractability theory**, which also deals with principle of *tokenization*. The question is split on elements that are appropriate in the incorporated lexicon. Attached to each item, there are three types of element: **relation, attribute** and **value**. These elements very well reflect the natural structure of language and allow the formulation of adequate query (Popescu, 2004).

4.5.4 Search engine

Finally after being rewrite into the artificial language the query is sent to the standard search engine (see chapter 3.2.2). These could be newly designed especially for the QA system, however, common search engines that exploit the traditional indexes are mostly used, as for example Google or Yahoo etc. (see chapter 3.2.2)

4.5.5 Answer extraction

As it has been mentioned earlier, QA systems vary significantly in the results presentation (see chapter 4.4.1). According to Clarke (2001) answer should be presented in the form of **values, names, phrases, sentences** or **brief text fragments** as the best way of results representation. Most of the research works focused on human perception recommend the maximum size of result´s data set about 50 kb which matches to the fragment representation. Larger data set is harder accepted by user and smaller is deficient, which results from the human cognitive capacities. (see chapter 6.3)

The format of full document, paragraph or large text fragment in response to the question is unlikely to be accepted by user. Even though the full document is not accepted as answer, it is often used as attachment to the particular result for extension of the user's view on the sought topic.

Currently, four different variations of result representation are normally used (Lin, 2003):

- **EXACT ANSWER** – only exact answer is returned. They are most often named entities, e.g., dates, locations, names.

- **ANSWER-IN-SENTENCE** – exact answer is returned with sentence from which the answer was extracted.
- **ANSWER-IN-PARAGRAPH** – exact answer is returned with paragraph from which the answer was extracted.
- **ANSWER-IN-DOCUMENT** – exact answer is returned with the full document from which was the answer extracted.

According to Lin (2003) the most comfortable way to present results is the **answer-in-paragraph**, which also supports one of the information retrieval theories that people mostly search with the purpose to learn something about the particular topic and not only to find the perfect match. This version presents enough information to satisfy the information need and also because the paragraph presents other additional contextual information.

The methods used for answer extraction vary. First, the full documents with potential answer are evaluated by weight algorithm (as for instance HITS or PageRank) already in the search engine. In the MULDER environment are these retrieved documents divided in chunks that likely contain the exact answer – these chunks are called *summaries*. MULDER ranks them and select the *N* best answer candidates. Afterwards, the system parses these summaries and obtains phrases of the expected answer type (Kwok, 2000). This approach is based on **bottom-up** method, where the answer is built up from the basic constituents into the full-value sentence based on the logic of natural language and belongs to the most common approaches.

4.5.5.1 N-gram method

The most popular method for answer composition is the **N-gram method**. All unigram, bigram and trigram word sequences are then extracted from returned results. Unigram is an n-gram which refers to the size 1 of implemented constituents of the meaningful utterance. All single existing words are unigrams. Bigrams refer to size of two and trigram to the size of three.

N-grams analyzed in the answers are scored according to their frequency of occurrence and the weight of the query that retrieved them. Afterwards, the candidates matching to the expected answer type are chosen and these subphrases – unigram, bigram and trigram – are tied together and reweight once more. From these results is extracted the one with the highest probability – according to the expected

answer – and the frequency of appearance which is also counted in the N-gram method (Azari, 2004).

It is clear, that the domain of answer extraction as well as the Query rewriting is large and challenging, but not impossible.

There is no doubt about the ingenuity of the system specially from the technical and systematic point of view. Whatsoever, the remaining question is on the user´s point of view and discussion on the capacity of the system development to actually reflects user´s information need and information processing.

5 Human information system

It is considered that usage of natural language as query language should simplify the users' access to information content via web-based search engines. The reason is to be seen in the modification of the search engines in the way of primary human information retrieval represented by thinking as the set of cognitive processes (Cejpek, 1998).

From the perspective of an information retrieval problematic, humans should be considered as a perfect information system. A decomposition of the human's information system will follow. Future research on artificial information retrieval systems should be held according to HIS.

At the present time, the understanding of HIS is far to be completely achieved. Clearly, knowledge about brain and neural system are still preliminary. However, there is few doubts that HIS stands beyond all human´s development. Although the outcomes of human being might not always be the best, it still shows better outcomes than any artificial information systems ever developed.

If the search techniques would reflect the users´ manners, it would be markedly easier to master them. Many studies are working on the idea of the exploitation of human cognitive processes in different ways, but not many of them justify the reason. It seems that these developments are based on assumptions without empirical or theoretical background in the field of human cognition.

This lack of empirical methods is based on the difficulties of cognitive research, where ethic and humanity forbid the possibility of experiments on healthy people. The experimental research is thus processed on people with brain disease, which brings limited data.

This chapter is focused on the way of human cognition, information processing and thinking from the physiological, philosophical and psychological point of view.

5.1 Neuropsychology

The following neuropsychological point of view is described briefly, because this topic belongs beyond the scope of this paper. However, it is probably the most

examined domain of HIS applied on artificial systems. It is understandable because of its coherency with computer-centered approach as the historically oldest approach.

The mainstream research in this field is based on the assumption that the centre of human´s information system is located in the brain. By the mean of brain functions this information system is able to perceive and react to the stimuli of the real world and affect the further behavior of humans in response to HIS.

However, the actual knowledge on neuropsychological processes remains significantly limited. Despite the development of few theories generally accepted by the experts; they remain isolated from accessible technologies and research methods.

Historically, mainstream researches on the brain functions have been mostly focused on individuals whose brain were either injured or affected by congenital cognitive defect. Clearly, such brain conditions triggers different answers than the average of human brain. Moreover, if the brain's structure and nervous system are similar for everyone, there are individual differences. Often, the issue of researches made on defected nervous system is the lack of knowledge on the state of the organ before the injury, therefore there are no relevant set of data to compare and support the implied hypothesis (Sternberg, 2009).

5.1.1 Human information system decomposition

The following description of nervous system is simplified for purpose of showing the relation with artificial information systems by showing it as the Human information system. The analysis of brain functionality is focused on the normal healthy brain as it emerges from the general accepted theories presented in Sternberg´s Cognitive psychology (2009) (Figure 6).

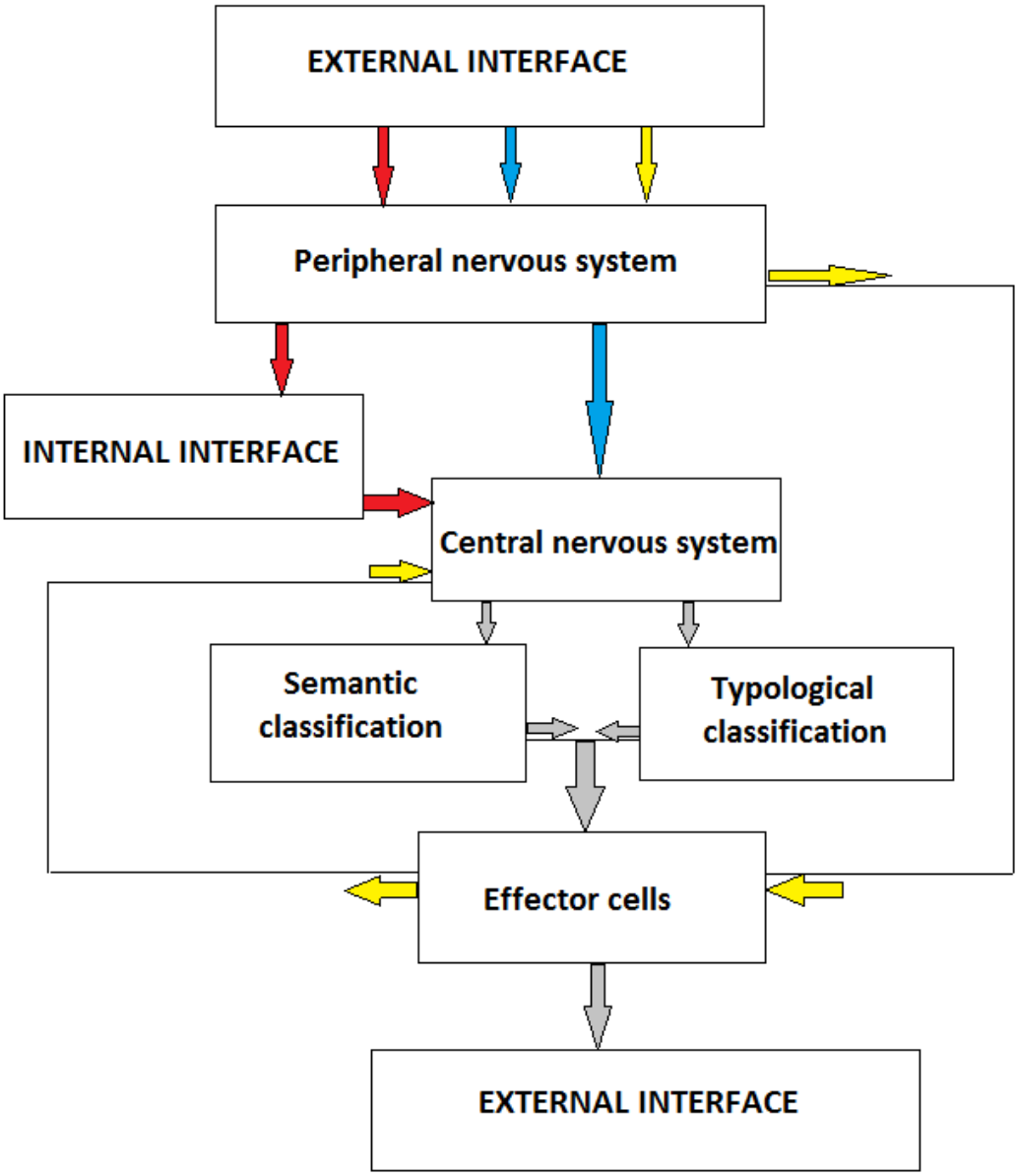

Figure 6 Human information system

5.1.2 Interface

The human information system has two types of interface: inner and external.

The surface of the human body works as the primary external interface of the system. The neural system acquires external impulses all over the body and transports them via neurons to the center of the system: the spinal cord and brain (Figure 7).

In general terms, the nervous system could be divided into **central** and **peripheral nervous system**. Even though they are interconnected and the central nervous system represented by brain is superior, they could work partly independently.

Internal interface transmits demands that did not originate as the reaction on external sensation. It is fully based on conscious inner cognitive processes and represents the second level of impulses evaluation (see chapter 6).

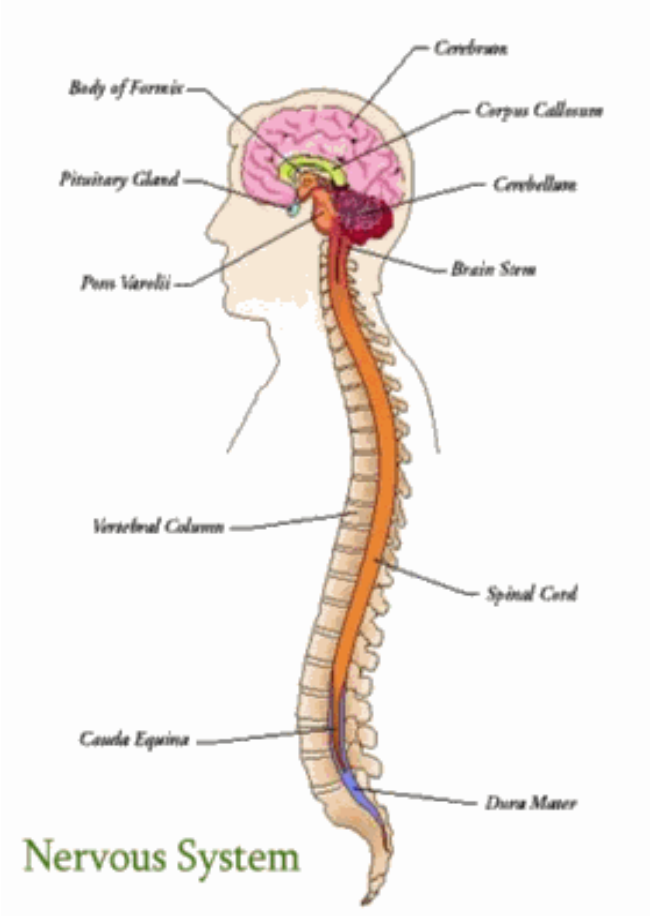

Figure 7 Central nervous system

5.1.3 Impulses

The communication between neurons is hold by electrochemical reactions that could inhibit or excite the impulse. The electrochemical impulse is the form of communicated information in the human information system.

As mentioned earlier, the communication has two modes: **inhabitation** and **excitation**. In the case of inhabitation, the impulse loses intensity and information ends up in the related center, where it had been send via the excited neurons. The excitation mode allows sending of the information from one neuron to another by raising its energetic level on the firing rate.

The information (impulse, signal) is stored in the brain according to the type of impulse. The general assumption is that this stored information is located in specialized storage areas that reflect their format: visual, olfactory, tactual etc. The position of these specific areas is still unknown; however their existence is proved, for example by the **split-brain syndrome** described by Michael Gazzaniga et al. (Sternberg, 2009)[10]

5.1.4 Neurons

There are three types of neurons: afferent neurons, efferent neurons and interneurons. Each type has different functionalities in the information sending process.

5.1.4.1 Afferent neurons

Afferent neurons (or sensory neurons) are situated in the interface of the human information system located on the body surface, where cells of receptors mediate the communication of the whole system with its environment. In other words they encode the external stimulus into internal electrochemical impulses readable for the nervous system.

[10] Split-brain syndrome is the chirurgical allocation of brain hemispheres. Specific nervous centers thus can not communicate between each other and the multidimensional picture of reality is limited by the one dimensional perception (Sternberg, 2009).

5.1.4.2 Efferent neurons

Efferent neurons or so called motor neurons are responsible for transmitting the impulse from the central nervous system to the effector cells responsible of the proper reaction of the whole human information system and the response to the external stimulus acquired by afferent neurons.

5.1.4.3 Interneurons

Interneurons are specific type of neurons that redirect the information to the right sphere of activity in the nervous system. Interneurons are the most important part of the whole system, because they are responsible for communication between afferent and efferent neurons.

Interneurons create an **association area** through which 75% of the brain capacity interferes and is located in the **grey matter**. This association area is responsible for process of thinking, conscious activities as planning and problem solving and the main function is the information storing in the form of memory.

5.1.5 Grey matter (cortex)

From the technical point of view the majority of cognitive processes is located in the grey matter, with its significant part providing memory processes. All cognitive processes are triggered by impulses of external or internal HIS interfaces.

The highest concentration of the grey matter is in the **cerebral cortex**, but it is located also in the **peripheral nervous system tangibly in the spinal cord**. This diversity and inherency of memory in peripheral nervous system enables HIS to respond to the acquired impulses with a fair degree of independence. For instance, the automatic processes and reflexes are possessed by the peripheral nervous system with consecutive redirection of the information to the central nervous system. It works like some kind of default options for the system responses that are proceeding unconsciously. This default option could by partly justify by learning and enabled by priming (see chapter 6.5).

Cerebral cortex was already mentioned in connection with the grey matter as the main center of memory and other cognitive processes. Another centre for these processes is the **basal ganglia** and **limbic system**. The most important part of the

limbic system for cognitive respectively processes is called **Hippocampus** (McClelland, 2009)**.** It is related to the learning processes and special type of episodic memory[11]. Hippocampus is responsible for creating new memories and for the retrieval of stored information in the memory. Moreover, it seems that the information classificatory function is also located in the hippocampus.

The last but not least important part of the brain related to cognitive processes is **cerebellum**. Its primary function is motor output, balance and posture. However it also takes considerable part in the creation of the unconscious procedural memory for learned and automatic activities.

[11] Episodic memory is special kind of explicit memory based on learned schemas and situational structures (Sternberg, 2009)

6 Cognitive processes

6.1 Cognition in general

Cognitive psychology is defined by Anderson (2000 ; In David, 2004) in the following quote as:

"Science concerned with the human mind, how it creates meaning, how it processes information it recieves (input) to develop responses (output), and how those responses (output) in turn can influence subsequent input."

The definition of cognitive science presented by Anderson is not complex. It is focused on the information perspective, and it does not say anything about other cognitive processes as, for instance, behavior or emotions[12]. Nevertheless, the aim of this paper is not to describe all of the cognitive processes, but only those closely related to the information processing and language exploitation, namely **perception**, **memory**, **language and problem solving**.

Furthermore, David (David, 2004) points on the fact that complex human behavior in the meaning of responses on stimuli is cognitively penetrable. The **cognitive penetrability** means that all human behavior is the response and direct output of human information system. Regardless on the character of the starting impulse, that could be conscious or unconscious, inner or external.

The second important assumption of this theory is the **permanent flexibility** and changeability of the human behavior regarding to the changeability of cognitive processes.

Nevertheless David (2004) points out, that some human responses are partially influenced by their genetic predispositions (Characters, Ego, etc.)[13].

---

[12] The problematic of emotional processes their influence on and information retrieval and information processing in general is largely elaborated by Kuhlthau (Kuhlthau, 2004). She brought the terms of information anxiety and the importance of inner feeling when information retrieving.

[13] This topic is beyond the scope of the paper, however genetic predispositions are the main topic of the classic psychologists as for instance C. G. Jung, S. Freud, Maslow.

Considering the searching memory and bringing context to bear in general information processing as natural human activity is still an intuitive assumption more than result of empirical research. However it is supported by empirical data mostly gained from controlled laboratory experiments. They do not describe the whole complexity of the cognitive process in their natural environment because of the artificial laboratory conditions (Hewett, 2005a,b).

6.2 Basic levels of cognitive processes

Basic cognitive processes can be classified according to different aspects. This introduction to the cognitive aspects of information retrieval presents some of them.

6.2.1 Conscious X unconscious model

One of the achievable options to divide cognitive processes is based on the existence of two different, but interlinked levels of cognition: **conscious** and **unconscious** (David, 2004).

Conscious level is responsible for the **explicit** cognitive processes (learning, memory, perception) and unconscious level for the **implicit** cognitive processing (for example favorable attitudes to the situations, evaluation processes etc.) (Greenwald, 1995).

6.2.2 Higher X lower model

Other possible distinction of cognitive processes is lower (surface) or higher (deep) character. **Lower** cognitive processes are easily and consciously accessible. **Higher** cognitive processes are albeit consciously accessible however difficult to recall (David, 2004).

Both levels participate to the information processing. However the higher level is more related to the information processing and thinking process.

6.2.3 Representation X computations

Another distinction of cognitive processes according to their function is claimed by David (2004). It is divided in two groups of representations and computations.

The **representations** embody the processes of real world by creation their reflection (see chapter 6.6). Representations refer to things, relations between them

as well as relations with their real likeness (e.g. word table is related to the real likeness of table – real existing table). One of the most important representational processes is the perception.

Computations refer to the processes of transformation. They operate with stored representations which are transformed from one to another in a rule governed manner to recreate the representation of the world. They proceed in the conscious as well as in the unconscious level. They embody higher cognitive processes of learning, problem solving, language processing etc.

6.3 Cognitive processes capacity

The capacity of the human nervous system processing is limited as well as the capacity of any other artificial system processor. This is the reason for application of **economy principles**, where the restricted capacity of reasoners´ working memory tends to work as little as possible (Manktelow, 1999). This principle is the basic aspect of thinking process (see chapter 6.5).

These limitations are represented by the value of the **channel capacity** for an absolute judgment. The capacity could differ from one user to another. Nevertheless it is hold by the nervous system in general range which is about **3,5 bit** of information. That is **7+/- 2** different stimuli that human information system is able to proceed to get some valuable response without crucial error rate in the judgment (Miller, 1956)[14].

The fact that visual based signals have larger channel capacity than the language processing is remarkable (Jacobson, et al., 1952). His validation is the multidimensionality of the language processing compared to the visual information processing that works only in one dimension.

6.4 Cognition and language

The cognition and language exploitation is generally accepted as being closely related to each other as one of the signs of human intelligence and difference

[14] The number 7+/- 2 is based on research held primarily by Pollack (Pollack, 1954) who detected the same channel capacity value in single or multi-dimensional stimuli experiments.

between humans and other species. Because of its lack of satisfying empirical research methods, this assumption is not satisfyingly provable. Most of researches on this topic are thus provided through theoretical abstract level.

Some scientists alleged to prove that thinking is language independent and that it exists also in simpler form of social cognition[15] between all animals as well as humans (Uhlíř, 2009). However, the predominant judgment claims that the difference in thinking is based on language mastering as one of the most complicated cognitive processes.

The significant changes in the language understanding and research brought for example Noam Chomsky and Wittgenstein in 1970s.

According to them (Susswein, 2009, Comsky, 2000) is phenomena of thinking and information processing involved mainly under the superior domain of cognition. The main language related cognitive processes are described in the following chapters.

6.4.1 Perception

"Perception is the process by which organisms interpret and organize sensation to produce a meaningful experience of the world. Sensation usually refers to the immediate, relatively unprocessed result of stimulation of sensory receptors in the eyes, ears, nose, tongue, or skin. Perception, on the other hand, better describes one's ultimate experience of the world and typically involves further processing of sensory input. In practice, sensation and perception are virtually impossible to separate, because they are part of one continuous process."

(Lindsay et al., 1977)

Perception is the starting point of all processes of human mind. As it arises from the upper mentioned definition, perception is the vein of people's world sensation. "World" entails in this case the individual environment of each person. Perception is related to the processes of meaning assignment to the acquired senses and their consecutive classification and organization.

[15]Social cognition is defined as sense of hierarchy and contextual understanding of situation and processes (Uhlíř, 2009)

In general terms perception allows people to acquire information about their particular reality and enables them to manage this reality by other cognitive processes to be able to exist in the particular environment.

The perception could be understood as the nonselective information acquisition on the level of external human information system interface (see chapter 5.1).

Unfortunately, perception is similarly as a majority of cognitive processes highly individual. The tudy of Herman A. Witkin (Sternberg, 2009) affirmed that individuals perceive the same situation in different ways. They see different details and contexts that are related to their knowledge base, experiences (higher cognitive processes) and partly also genetically predispositions.

For example, it is most likely the reason why witnesses of a car accident differ in its description. The sensation is directly proceeded by computational processes and reorganized. The difference of this consecutive processing is the cause of the different perception and sensation of the situation.

This theory is supported by results of optical illusions experiments. There were significant differences in visual perception can be observed for each individual. The aim of this test is to provide individuals with a single picture that has two different ways of interpretation (Figure 8). Most likely, a single individual saw only one way of possible representation until they are informed of the existence of second representation. By proceeding this new information, they were usually able to recognize both variations (Sternberg, 2009).

Figure 8 Optical illusion

Other important phenomenon of perception is the fact that the perception is different of what is actually seen. This interesting characteristic is caused by the function of sensorial memory which makes our mind to think that what we percept is actually what we see (Sperling, 1960) (see chapter 6.4.3). We tend to trust our mind more (literally sensorial memory) than the reality and thus we project representations from sensorial memory in the reality until 150 ms (Sternberg, 2009).

6.4.2 Natural language and perception

Natural language has a lot in common with perception. It is the only way how to acquire language in any form. There are three overbearing ways of language perception: **visual**, **acoustic** and **tactual**. The focus of this paper is on the visual language perception as the most frequented way of communication in the web environment.

The basic process of visual language acquisition is **reading**. It is one of the most complicated processes arched over by cognitive psychology. Nevertheless it is gateway for textual information processing.

6.4.2.1 Model of language processing

According to theories published by Lashley (1951), Chomsky (1957) and Garrett (1957) there were two levels model of language processing. This model consists of two main separated frames of information processing: **lexical** and **semantic** (Dell, 1993).

Model considers the serial processing from the lower to higher cognitive activities when acquiring and inverse direction for language processing. Lexical frame is thus included in the lower cognitive activities. The structure of lexical frame consists of three sublevels elaborating the incoming visual signal on the level of **shape**, **letter**, and **word**.

In general, the function of lexical frame is defined as the ability to understand the text along the formal representation. The interesting finding is, although the whole model works with serial processing sequences, in the lexical frame as well as in the semantic frame where the processes are exploited collaterally and contextually.

These findings are based on the research held by James McKeen and Ray Cattell from 1886. They claimed that people could recall more letters from words than meaningless strings of letters (Grainger, 2003; In Sternberg, 2009), as well as the ability to read faster the meaningful text (on lexical level) than randomly written letters. The results of these experiments stand beyond the theory of **Word superiority effect**.

This is based on the contextual relations between sublevels of the lexical frame. These relations are most probably based on prototypes categorization (see later). And it takes in account the visual and lexical level of the text (morphology, phonetics and phonology).

Cattel's theory was later extended in the **Sentence superiority effect** theory that belongs between the lexical and semantic frame (Cattell, 2006). According to the studies this theory belongs rather in the semantic frame, while the contextual information, which is important for syntax is based more significantly on the semantic relations than in the case of lexical frame processes.

6.4.3 Memory

Memory processes are other important processes related to the information processing of human information system (HIS). Its broad function is to store the information or knowledge and relations between them.

As well as the channel capacity for perception is limited, there are also limitations for memory as the information storage. These limits are highly individual and extendable by particular rehearsal (Hewett, 2005).

Historically, the structure of memory is described as system of three independent spheres of **sensorial memory** (SM), **short term memory** (STM) and **long term memory** (LTM). All kinds of memory are located in the central nervous system represented mainly by brain.

The statement of this historical theory underlines that all kinds of memory work independently and serially according to different manner. The key idea of this theory, for the purposes of this paper, is the difference of information classifications in particular memory levels (Sternberg, 2009).

6.4.3.1 Sensorial memory

SM is the primary storage for most of the acquired (*perceived*) information. Sometimes is SM involved under the STM (Sperling, 1960). However its importance and purpose is significant enough to be described as single unit.

SM is represented by **iconic memory**. Its capacity is about 4 -5 chunks of information, which is not enough to create any judgment about the perceived situation as presented before (see Cognitive process capacity). The duration of the display in the iconic memory is approximately 250 ms after the offset of an impulse (Sperling, 1960).

The character of the SM representation is predominantly visual. It has the character of the **icons** as the reflection of reality, without any intervention of classification or organizational structure of further memory phases.

Beside the visual SM representation also exist other variations of SM according to the form of sensation; nevertheless the visual sensorial memory is more fully elaborated. And, as it has been mentioned earlier, the process of reading as the language processing is primarily based on the visual perception. According to this statement text is primary processed as the pure visual information in the iconic SM.

Information acquired via SM is forwarded to the STM and later to the LTM, where information is processed in the contextual environment of the organizational structure of the memory and reorganized in the knowledge base.

6.4.3.2 Short term memory

STM is understood as the second level of information processing in the central nervous system. The standard capacity of STM is the same as the channel capacity for absolute judgment (see chapter 6.3). As well as could be percept 7+/- 2 impulses to make an absolute judgment in the HIS, it could be stored 7+/- 2 chunks of information in STM.

The duration of the data holding is limited from 12 – 30 seconds. This limitation is extendable by maintenance rehearsal in the meaning of simple repetition of impulses. (Hewett, 2005)

This capacity enables STM to make a judgment and thus react on the initiating impulse of information process coming from SM. However, because of the duration limitation of STM the response is mostly automatic and often unconscious (Sternberg, 2009).

The representation in the STM is assumed analyzed on two levels: syntactic and semantic. Both levels are represented by a single possible way of expression without the direct connectivity to other possible structures and schemas nor the LTM level.

This **single-structure** processing is based on the priming process as the primary classification (see chapter 6.5).

The information stored in the STM is consequently transmitted in the LTM where single-structure classifications are implemented in the likely multidimensional semantic structure.

The scientific understanding of the STM function and processing is changing and the current approach of cognitive psychology claims that the STM is rather part of the LTM than independent space for information storage. This modern understanding of STM is though implicated in the operational memory.

6.4.3.3 Operational memory

Operational memory (OM) or **working memory** represents STM as the part of the whole memorizing processes exploited for the actual cognitive processing. Processes in this type of memory are largely automated and fall into the unconscious level of memory (see chapter 6.5).

These automated processes help to master daily life in the meaning of iterative actions or situations that are later applied without deep conscious mental activity. On the other hand the very same unconscious automated behavior could make it difficult to apply new or change the old already automated processes. This block could cause serious problems in daily life because of errors and mistakes in the situational evaluation. However, it also slows down the learning process and limits the creative problem solving.

The depending part of the memory process underlies OM to the same classifications of information as the LTM. And the difference between them appears

as the difference between conscious and unconscious information processing see LTM).

6.4.3.4 Long term memory

LTM is the human information storage with large capacity for long period of time. Although there is an obvious existence of capacity limitations, there is no empirical base for its measurement. Nevertheless, its capacity is not static, but extendable as well as the capacity of STM. Because of the different structure and function, the rehearsal activities for LTM capacity extension are based on the elaborative principles (Hewett, 2005). That implies to focus on the restructuring and reorganization of information classification in the storage rather than memorizing the single semantic representation stored already in the STM.

The LTM stores several types of information including such as motor and perceptual skills, knowledge and the entire representation of the world.

The modern approach represented by Anderson (2004) claims that the classification structure of LTM influences the information classification structure in STM. He assumes that STM has even the same structure than the LTM, and the difference in responses is based on the different time duration of the information processing (Sternberg, 2009).

LTM, according to the modern concept, process all information incoming from the SM on two levels. The first on the unconscious level defined traditionally as STM. Afterwards on conscious level commonly understood as the LTM. Nevertheless the difference is in the volume of appreciation of the submission of acquired information in the inner classification structure.

Regardless of the approach, both hypothesis claim, that the LTM is the center of the syntactic and semantic classification schema. Characters of this schema are: multidimensionality, connectivity; dynamicity and heterogeneity (see chapter 6.6).

6.5 Priming

"Priming is process of transporting the unconscious structures related to the current acquired stimuli in the conscious state of mind.!

(Sternberg, 2009; p.91).

Priming was already mentioned as process of STM as the very first analyzing of acquired information. The matter of this process is dual.

The first matter of priming is **physiological**. This process stimulates the electrochemical impulse which enables the communication between the particular neurons. The excitation mode allows sending the information from one neuron to another by raising its energetic level. Impulse is send in more than one way. However there is constantly one (in some cases more than one) way which drowns other signals by its intensity and assumes the leading status of conscious information processing. Other paths are also inhibited and being related, they are suppressed in the unconscious.

The second matter of this process is abstract on the level of the primary information retrieval: **thinking** (Cejpek,1998). This abstract level has two basic options of processing.

The first common option is based on deficiencies in the inner representational world classification. User cannot unambiguously evaluate the meaning of the incoming information, because it does not sufficiently match to any present representation.

In that case he activates more than one semantic structure related to the meaning of the information. According to this process he gets broader knowledge about the acquired information. He reconstructs the current knowledge about the topic and complements the new information in the current structures. If the result is still not sufficient and the information need is deficient, he require for more coherent information according to the current knowledge base. This bottom up process represents the **constructive** character (Sternberg, 2009; p.231) of the information retrieval which is influenced by the current state of mind. The whole process

proceeds on the unconscious level and its running can be determined by the emotions (feelings of anxiety, uncertainty, doubt and confusion, etc.) (Kuhlthau, 2004).

The second option of priming process is based on the existing appropriate representation in the inner semantic structure. User is familiar with the meaning of the information and he activates unconsciously only one related structure. The other possible consequences are inhibited and less accessible, because they are overlapped and drown by the usual most common meaning in the inner semantic structure. This character closes and makes difficult the further development of the possible understanding the situation in different way and the constructive character of new information retrieval is poorly applicable. This means of priming decreases the plausibility of the creative problem solving and blocks the formation of new connections in the inner semantic structure and mental models.

6.6 Representation

The term *representation* has been mention previously in this study. Representation is the main entity of information processing on the abstract level.

The main characteristic of cognitive processes is the availability of humans to understand and describe (*represent*) objects, situations, processes and the world in general in a way to create and share these representations with other humans.

6.6.1 Knowledge and representation

At the end of the 20th century has been established the theory of a society which is economically based on information, its usage and its distribution (Porat, 1977).

At the beginning of the 21st century emerged theory of knowledge society, which is a modification of the previous information society theory in the manner of substitution of knowledge in behalf of information. Where the main difference between information and knowledge is in contextual relations.

Information in general is a contextually independent unit, which is indexable and organizable according to norms, standards and it is independent from its author.

Knowledge is on the other hand based on contextual engaging. It means that knowledge can be defined as „information in use“ which is involved by experiences of author and specific environment where the knowledge is developed. Knowledge is

not necessarily expressed. Its character is human based. It takes place in human minds, where the inner knowledge classification helps to understand and manage the human interaction with reality.

It seems, that the knowledge is in fact the pragmatical reflection of information presented by **intellectual capital** of individuals (Bukh, 2001). Knowledge is thus the basic matter of the abstract level of information processing.

It is evident that representation and knowledge definition have a lot in common. Some scientists consider representation and knowledge almost as synonymous where representation emphasizes the format in which is knowledge stored (Susswein, 2009).

According to Wittgenstein **representational cognition** is understood as solely human characteristic (Susswein, 2009). Where representations substitute the real objects and model the reflection of the real world on the abstract level. Thus, human understanding of the real world is based on their inner individual knowledge schema as the representation of the outer real world. These schemas refer to a cluster[16] of knowledge that contains information about core concepts, the relations between these concepts and knowledge about how and when to use these concepts (Chinnappan, 1998).

The individuality and sophistication degree of such schema is based on socio-cognitive aspects of personal's environment (see chapter socio-cog approach).

Knowledge schemas are in traditional psychological literature defined as the **mental models** (Chinnappan, 1998).

[16] Clustering theory in general is a specific method of data classification which is broadly exploited in IRS, where clusters are not predefined and their emergence is based on probabilities methods, based on the information space theory. However there is plenty of models described for example by Osmar Zaïane (1999).

6.6.1.1 Mental models

Mental models according to M. David Merrill´s (2000) Component Design Theory (CDT) consist of two major components:

1. **Knowledge schema** - knowledge organizational structure
2. **Mental operations** - processes using this knowledge

Knowledge as the basic matter of the mental model could be represented by four different types of knowledge objects:

1. **Entities** – representing things (*objects*)
2. **Actions** – procedures that can be applied on, to or with entities or their parts.
3. **Processes** –represent events that occur often as a result of some action. It is knowledge about how things work.
4. **Properties** – attempt the qualitative or quantitative value of the other knowledge objects.

This theory is well adapted to any content to be taught. And it significantly reflects the human unique ability to conceptualize or to place entities, actions, and processes into categories.

The general concept of the knowledge processing in the CDT is shown on the following diagram. Where is reflected the network of basic knowledge objects and their relations called PEAnet (Process, Entity, Activity Network) are supposed to be applied on any information processing (Figure 9).

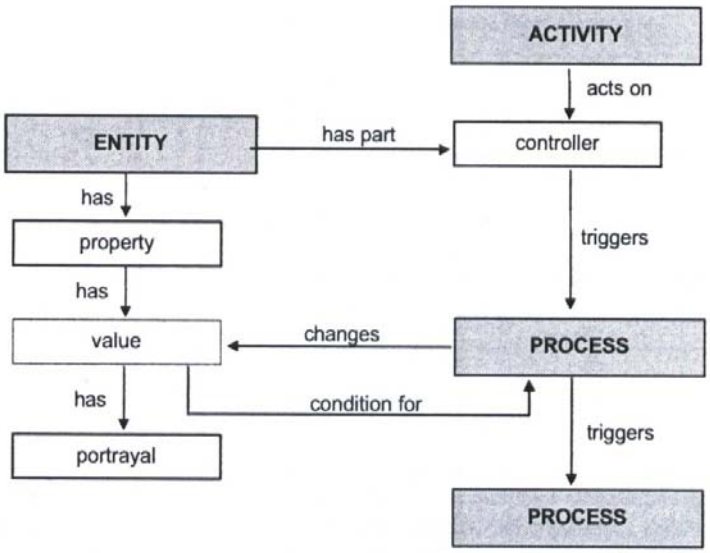

Figure 9 Casual network process knowledge structure (Merill, 2009)

Mental models are defined as cognitive representations that individuals construct during various learning situations. The process of modeling is based on aligning the current solved situation anomalies with the components of the existing knowledge

6.6.2 Inner language

One of the most often employed code system for cognitive processes is natural language. Nevertheless the individual differences in social and cognitive background enables emergence of **inner language** or so called **language of thoughts**. The language system is thus one of the code systems for information storage. This code serves as the main supportive code system for knowledge schema because of its general accepted rules (see Language).

Given the influences of the outer background language of thought can be incoherent with the general accepted system of natural language. In other words: everyone has his own language or **sublanguage**. The existence of sublanguage supports the idea that users have some common general language regulations, but in the mind's operation of such regulations are independently reshaped by individual contexts (Chinnappan, 1998).

6.6.3 Symbols

The form of representations is established as symbolic, where symbols could be strings of letters as well as image or other substituting percept.

It appears that language as the code system is limiting for inner representations in the way of expression. For instance some of the thoughts and representations are not possible to express by language. User is feeling confused, however, he knows that the representation of reality exist, he understands the situation, but he miss the expressions to describe it in the way which would be understandable for others. That is the effect of inner individual languages; it has meaning only for individuals.

Because of the language code deficiency, the representation must be completed by other supportive code system, mostly represented by visual perception. Representation consists of more than one descriptive code. They communally depict the reflection of reality.

6.6.4 Information space

As it has been mentioned earlier, the reflection of reality has to be contextual multidimensional anchored. The current approach is based on the creation of contextual or semantic network. This network can be defined as 2D or 3D information space that fills these demands.

Information space is defined as the "*set of relations among items held by an information system"* (Ingwersen, 1996).

Information space applied on human cognition consists of symbols (representations) and relations among them set by user´s knowledge (Sabol, 2002).

The information space theory is beyond most of the IR systems even though it is hidden behind the one dimensional representational level.

6.6.5 Meaning blindness

Representations in their raw matter do have nothing to do with semantic meaning.
(Proudfoot, 2009)

„*Understanding a story is not a mere process of identifying truth conditions of a series of sentences, but is a construction process of building several partial models such as a model of the environment in which the story takes place, a model of mental attitudes for each character and a model of the verbal interactions taking place in the story.* „

Moulin Bernard (1998)

It reflects Wittgenstein's **meaning-blindness** theory, where the inner representations have any meaning (Proudfoot, 2009) before their connection and combination with the real world and other representations. In other words: the representation acquires semantic meaning after the transport in the knowledge structure by applying the process of cognition and transformation into the knowledge (Figure 10).

Therefore, according to Wittgenstein, knowledge is understood as assimilated mental model from the conscious to the subconscious level of cognition processes, that is presented by knowledge schema as the picture of reality. The knowledge schema is used to understand the real world.

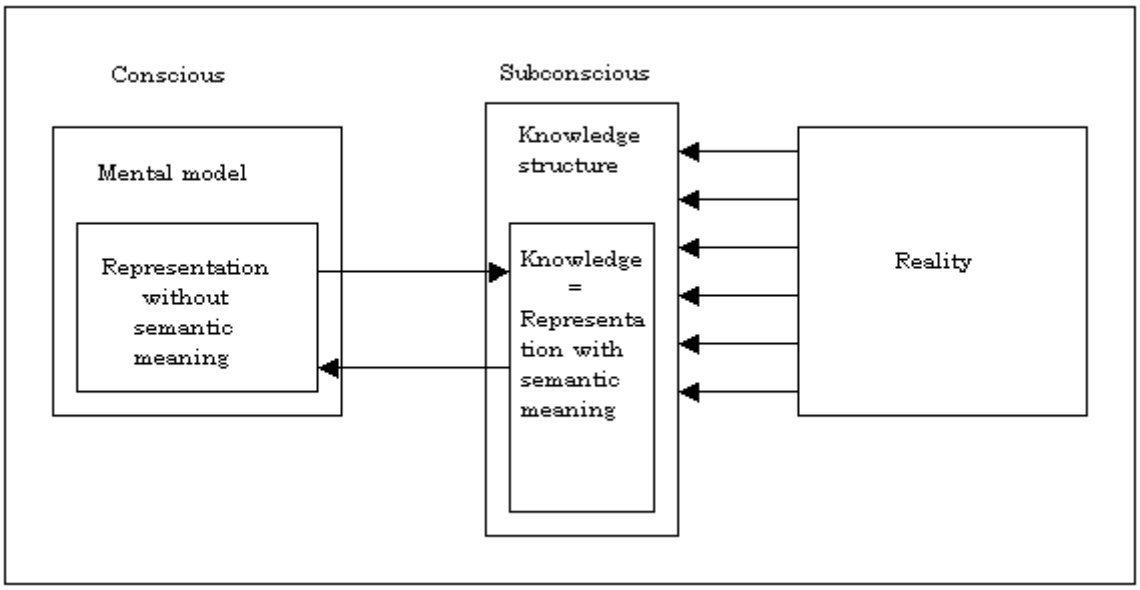

Figure 10 System of cognition according to Wittgenstein

7 Conclusion

As presented, users should be the moving element of all research in domain of information retrieval. The reason is simple and doubtless. Users stand at the beginning and end of the information retrieval process. There is no need for any research on this hypothesis, because IRS domain is human domain developed by users for users. As the purpose of this domain is to enable users to find relevant information faster and easier, it is necessary to focus on them and their needs.

This research was presented the problematic of information retrieval in the web environment and its focus was given on the special kind of IRS - the natural language question answering systems.

The topic of this thesis is certainly large. Its scope is given mainly by its multidimensionality. The aim of this paper was not to span the whole problematic and bring a perfect solution. It was rather to bring the general overview of the problematic with basic applied solutions. To a certain extent, the lack of multidimensional knowledge is alarming.

While searching for this paper, it has been surprisingly difficult to find information focused on the aim of the developing such systems. Most of the articles found were presenting the aim as the development of system which is able to answer on questions performed in natural language. However, I did not find any study about the user approach to this problematic, which would defend the necessity of developing such systems.

This problematic is strongly connected to the cognitive psychology and linguistic, where these topics are discussed, nevertheless again without multidimensional interlink.

This topic is more adequately elaborated in linguistic domain that is already significantly applied in the information studies and information technologies, artificial language and its relation to cognitive science is also well defined (Cattell, 2006).

To a certain extent, the cognitive psychology is still neglected as the part of this problematic and although the processes in nervous system are deeply studied, the results are rarely interlinked with other domains. Cognitive aspects are mostly replaced by information studies often examining user's attitude from the position of his interaction with the device (device centered), than the influence of the device on users attitude and its usability.

The assumptions seem obvious n the case of natural language exploitation in web search engines:

Natural language search engines will lead to the faster and more accurately information retrieval, just because everyone use natural language to communication and it is much easier to express information need in natural language, which is deep rooted in user´s mind and he need not to learn its usage.

However information need is the stumbling block, of the whole research.

As it has been mentioned earlier, users do not always know, what they want to know respectively, what they want to find (Spink, 2001). This behavior belongs to human predispositions cased by limited cognitive processes. It means that they will not be able to express their exact information need in any language, not even in their native natural language if they do not know what they want.

However, natural language is automated and deep rooted in human mind as communication tool. It still represents complex and one of the most complicated human cognitive processes.

From the external point of view natural language (one concrete language, we do not take in account multilingual problems) inconsistent and the data volume produced by people is uncontainable to achieve the reasonable and final understanding of its structure and use (see chapter 4). These external attitudes of research are based on objective contemplation and so far results of this research are satisfactorily adapted in other domains. However language studies also have internal point of view. It is focused on human as the language producer and this research is held on cognitive individual level. This approach contrary to the external attitude is highly subjective. The results show the similar system of language providing on the level of syntactic and semantic language structures, however the deeper schemas of these structures are based on individual experiences and knowledge influenced by

different individual environment. The internal attitude is empirically hardly achievable and its research is focused mainly on theoretical and philosophical basis. Its application in other domains is thus more on experimental than empirical level of research. Natural language web search engines are then a perfect example of such experimental system.

Natural language search engines are close to encyclopedic systems based on the perfect match strategy as the most common strategy of information retrieval. The whole process of natural language queries rewriting and the answer extraction is surprisingly well managed, nevertheless the results are not perfect. The main problems are in the query rewriting caused by lack of knowledge about internal natural language processing.

If the basic hypothesis considers natural language search engines helping user to express its information need, it contrast him against the device in the communication partner position. This position highly influence the way of expression and it burdens user by additional cognitive processes that are usually not used in work with standard web search engines based on conceptual, textual level communication.

As presented in chapter 6, human language perception is running simultaneously on two levels: lexical and semantic. During the thinking process the lexical level is inhibited and the process of thinking is running on conceptual level based on multidimensional information format. The inhibited lexical level of natural language processing is still presented on the simplified level. That is proved by sentence superiority effect, which enables user to understand the lexical meaning of the sentence (in the way of syntactic understanding) even though it has not semantically meaning. However the lexical level is in the inner communication exploited very marginally.

In the case of natural language search engine, the lexical level is excited in the meaning of the question answering search engine. Lexical level then provides the template for the information need formulation, in the case of semantic expression. This formulation is based on semantic information encoding sufficient to the socially acceptable forms of language communication.

In the light of these facts, this type of natural language search engines does not provide the simplified way of information retrieval from the individual cognitive

(language processing) point of view. This conclusion supports the third hypothesis of this paper representing the idea of multidimensional and format heterogenic human information processing (acquisition).

The same hypothesis claims that the visual approach would be preferable before other approaches. Nevertheless from the text results that any textual whether conceptual or fulltextual approach represents visual approach. It arises from the fact that web based textual information processing is provided by reading (presented in chapter 6.4.1).

The first related hypothesis was not completely proved. However there were theories of human information classification in memory based presented on icons and symbols, which again supports the idea of visual information processing and acquisition.

One of the theories related to this topic which has been presented earlier was the theory of human knowledge multidimensional clustering method as method for information storage. That could be promising theory for further information search engine development. It is namely closely related to the information clustering theory which is nowadays usually used in the systematic part of search engines, and thus already well adapted on information environment. The upgrading would be the permeation of this theory with the interface and the user part of the system. By applying visual approach on this clustering method may originate better organizational tool for the overflow of the potentially right answers.

7.1 Further work

There is no doubt about the importance of user standing behind all processes related to information retrieval. And even if the socio-cognitive approach is fully reasonable, the very beginning of all processes lies in the user itself in his individuality that decided to find the answer on some more or less exact question.

The necessity of making information retrieval systems easier to use is obvious. As it has been said, information retrieval systems are systems developed by people for people. That is why the development should leads towards the user´s thinking processes, because the thinking *is* information retrieval (Cejpek, 1998).

Neuropsychology is significantly applied in information domains. Nevertheless there appeared interesting topic of HIS based on two level interfaces enabling human to cope with two dimensions of information research: speed and deepness.

These two dimensions are resolved in HIS by two different interfaces that also have different functionality and expletory each other. The development of such system which would apply two different algorithms for the retrieval process to find the fastest and the precise answer would be probably a significant progress in the research. This would fully correspond with the HIS.

In the ideal case, the whole interface would be based on visual approach by using the conceptual text and depicted relations between them based on dynamic clustering method.

HIS is not a perfect system; nevertheless it is not perfect, because we do not understand it. As we are going further with our research as the closer we are to understand of human as the information system.

8 Bibliography

ALLEN, Bryce. *Information Tasks : Toward a User-Centered Approach to Information Systems*. Orlando : American Press, 2000. 308 s.

ANTWORTH, Evan, L. *User's Guide to PC-KIMMO : Version 2* [online]. [s.l.] : [s.n.], c1995-2000, 2000 [cit. 2010-07-29]. Dostupné z WWW: <http://www.sil.org/pckimmo/v2/doc/guide.html>.

AZARI, David, et al. 2004. *Microsoft Research* [online]. Microsoft, 2004 [cit. 2010-07-29]. Web-Based Question Answering: A Decision-Making Perspective. Dostupné z WWW: < http://research.microsoft.com/en-us/um/people/sdumais/uai2003-qa-dt.pdf >.

BALÍKOVÁ, Marie. c2009. Natural language. In *KTD : Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online databáze]. Praha : Národní knihovna České republiky, 2009- [cit. 2010-07-29]. Systém. č.: 000001618. Dostupné z WWW: <http://sigma.nkp.cz/cze/ktd>.

BATLEY, Sue. 2007. The I in information architecture : the challenge of content management. *Aslib Proceedings : New Information Perspectives* [online]. 2007, vol. 59, no. 2, s. 139-151. [cit. 2010-07-29] Dostupný také z WWW: < http://www.emeraldinsight.com/0001-253X.htm >.

BAWDEN, David. 2006. Users, user studies and human information behaviour : A three-decade perspective on Tom Wilson´s "On user studies and information needs". *Journal of Documentation* [online]. 2006, vol. 62, no. 6, s. 671-679. [cit. 2010-07-29] Dostupný také z WWW: < http://www.emeraldinsight.com/0022-0418 >.

BELKIN, N. J., et al. 2000. Support for Question-Answering in Interactive Information Retrieval : Rutgers´ TREC-9 Interactive Track Experience. In *Proceedings of TREC-9*[online]. New Brunswick : [s.n.], 2000. s. 1-12. [cit. 2010-07-29] Dostupné z WWW: < http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.8638 >.

BERENDT, Bettina; KRALISCH, Anett. 2009. A user-centric approach to identifying best deployment strategies for language tools : the impact of content and access language on Web user behaviour and attitudes. *Information Retrieval*. 2009, vol. 12, no. 3, s. 380-399.

BORGMAN, Christine L. (1989). All Users of Information Retrieval Systems are Not Created Equal: An Exploration into Individual Differences. *Information Processing and Management*, vol. 25, no.3, s. 237–251.

BRAND-GRUWEL, Saskia; WOPEREIS, Iwan; WALRAVEN, Amber. 2009. A descriptive model of information problem solving while using internet.

Computers & Education [online]. 2009, vol. 53, no. 4, s. 1207-1217. [cit. 2010-07-29] Dostupný také z WWW: < http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VCJ-4WSWYDN-1&_user=10&_coverDate=12%2F31%2F2009&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_searchStrId=1415676936&_rerunOrigin=scholar.google&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=1978d16c64ecaf264c02c0bc05ccb067 >.

BREJCHA, Jan. 2009. Co skrývá uživatelské rozhraní?. In *Uživatelsky přívětivá rozhraní*. Ed. Alena Červenková, Michal Hořava. Praha : Horava & Associates, 2009. s. 43-52. ISBN 978-80-254-5295-0.

BRYS, Catherine M. 2004. *Discussion paper on content management strategy*. Glasgow, 2004. 12 s. Seminární práce. University of Glasgow.

BUKH, P. N.; LARSEN, H.T.; MOURITSEN, J. 2001. Constructing intellectual capital statements. *Scandinavian Journal of Management*. 2001, no. 17, s. 87-108.

CASE, Donald O. 2006. Informatrion Behavior. In *Annual Review of Information Science and Technology*. Cronin Blaise. [s.l.] : [s.n.], 2006. s. 293-327.

CATTELL, Ray. 2006. *An Introduction to Mind, Consciousness and Language*. London : Continuum, 2006. 226 s. ISBN 0-8264-5516-6.

CEJPEK, Jiří. 1998. *Informace, komunikace a myšlení.* Praha : Karolinum, 1998. 179 s. ISBN 80-7184-767-4.

CHARNIAK, Eugen. 2000. A Maximum-Entropy-Inspired Parser. In *Proceeding of NAACL-2000*. Providence : [s.n.], 2000. s. 132-139.

CHENG, Lisa Lai Shen . 1991. *On the typology of Wh-questions***.** [online]. Massachusett : MIT Press, 1991. 229 s. Dizertační práce. Massachusetts Institute of Technology. Dostupné z WWW: <http://dspace.mit.edu/bitstream/handle/1721.1/13938/24960288.pdf?sequence=1>. [e-akademická práce].

CHINNAPPAN, Mohan. 1998. Schemas and mental models in geometry problem solving. *Educational Studies in Mathematics*. 1998, 36, s. 201-217.

CHOMSKY, Noam. 2000. *Language and thought*. 5th edition. [s.l.] : Moyer Bell, 2000. 96 s. ISBN 1-55921-076-1.

CHOWDHURY, Gobinda G. 2004. *Introduction to modern information retrieval.* 2nd edition. London : Facet, 2004. 474 s. ISBN 9781856044806.

CHOWDHURY, Gobinda G.; CHOWDHURY, Sudatta. 2001. *Information sources and searching on the World Wide Web*. London : Library association publishing , 2001. 174 s.

CLARKE, Charles L. A.; CORMAK, Gordon V.; LYNAM, Thomas R. 2001. Exploiting Redundancy in Question Answering. In *SIGIR´01*. New Orleans (Louisiana) : [s.n.], 2001. s. 358-365.

CLINE, Ben E.; NUTTER, Terry J. 1990. Implications of natural categories for natural language generation. In *Current Trends in SNePS — Semantic Network Processing System* [online]. Berlin : Springer, 1990. s. 153-162. [cit. 2010-07-29] Dostupné z WWW: < http://www.springerlink.com/content/y325043w6g2522u5/ >.

CRUDGE, Sarah E.; JOHNSON, Frances C. 2004. Using the Information Seeker to Elicit Construct Models for Search Engine Evaluation. *Journal of the American Society for Information Science and Technology*. 2004, vol. 55, no. 9, s. 794-806.

DAVID, Daniel; MICLEA, Mircea; OPRE, Adrian. 2004. The information Processing Approach to the Human Mind : Basics and Beyond. *Journal of Clinical Psychology*. 2004, vol. 60, no. 4, s. 353-368.

DCMI Home : Dublin Core Metadata Initiative. c2010 [online]. c2010 [cit. 2010-07-29]. Dostupné z WWW: <http://dublincore.org/>.

DELANNOY, Jean-Francois. 2001. What are the points? What are the stances? : Decanting for question-driven retrieval and executive summarization. In *Annual Meeting of the ACL : Proceedings of the workshop on Human Language Technology and Knowledge Management* [online]. Canada : [s.n.], 2001. s. 1-8. [cit. 2010-07-29] Dostupné z WWW: < http://portal.acm.org/citation.cfm?id=1118232 >.

DEY, Anind K. 2001. Understanding and using context. *Personal and Umbiquitous Computing*. 2001, vol. 5, no. 1, s. 4-7.

Domain Counts & Internet Statistics. c2010. [online]. c2010 [cit. 2010-07-29]. Dostupné z WWW: <http://www.domaintools.com/internet-statistics/>.

DUMAIS, Susan, et al. 2002. Web Question Answering : Is More Always Better?. In *SIGIR´02* [online]. Tampere (Finland) : [s.n.], 2002. s. 291-298. [cit. 2010-07-29] Dostupné z WWW: < http://research.microsoft.com/en-us/um/people/sdumais/SIGIR2002-QARevised.pdf >.

FORD, Nigel. 2005. *New directions in Cognitive Information Retrieval*. Ed. Amanda Spink, Charles Cole. Netherlands : Springer, 2005. New cognitive directions, s. 81-96.

GENTILE, Anna Lisa, et al. 2008. Lexical and Semantic Resources for NLP : From Words to Meanings. In . I. Lovrek, R.J. Howlett and L.C. Jain. Berlin : Springer, 2008. s. 277-284.

GLEZERMAN, Tatyana B. ; BALKOSKI, Victoria I. 2002. *Language, Thought, and the Brain* [online]. [s.l.] : Springer US, 2002. Basic Factors in the Human

Brain's Differentiation Underlying Cerebral Organization of Language Ability , s. 3-25. [cit. 2010-07-29] Dostupné z WWW: < http://www.springerlink.com/content/xt10117384823177/ >.

GOH, Ong Sing, et al. 2007. A Multilevel Natural Language Query Approach for Conversational Agent Systems. *IAENG : International Journal of Computer Science* [online]. 2007, vol. 33, no. 1, s. 1-7. [cit. 2010-07-29] Dostupný také z WWW: < http://ainibot.org/index.html#DP >.

HAVLÍČKOVÁ, Klára. 2008. *Teoretické základy selekčních jazyků*. Praha, 2008. 25 s. Referát. Charles University in Prague.

HEWETT, Thomas T. 2005a. Designing with Human Memory in Mind. In *Mobile HCI´05*. Salzburg (Austria) : [s.n.], 2005. s. 363-364.

HEWETT, Thomas T. 2005b. Informing the design of computer-based environments to support creativity. *International Journal of Human-Computer Studies* [online]. 2005, vol. 63, no. 4-5, s. 383-409. [cit. 2010-07-29] Dostupný také z WWW: < http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WGR-4G5BJW8-1&_user=10&_coverDate=10%2F31%2F2005&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_searchStrId=1415673156&_rerunOrigin=scholar.google&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=c77f56861c8d1fffd4efd1c2acb8bcd3 >.

HICKL, A. 2008. Answering questions with authority . In *CIKM´08*. California : [s.n.], 2008. s. 1261-1270.

HICKL, Andrew. 2008. Answering Questions with Authority. In *CIKM´08* . Napa Valey (California) : [s.n.], 2008. s. 1261-1270.

HJORLAND, Birger. 1998. Information Retrieval, Text Composition, and Semantics. *Knowledge Organization*. 1998, vol. 25, no. 1/2, s. 16-31.

INGWERSEN, Peter. 1996. Cognitive Perspectives of Information Retrieval Interaction : Elements of a cognitive IR Theory. *Journal of Documentation*. March 1996, vol. 52, no. 1, s. pp. 3-50.

IOANNIDIS, Yannis. 2008. From Databases to Natural Language : The Unusual Direction. In *NLDB 2008*. E. Kapetanios, V. Sugumaran, M. Spiliopoulou. Berlin : Springer, 2008. s. 12-16.

JAKOBSON, R., Fant, C. G. M., & Halle, M. 1952. *Preliminaries to speech analysis*. Cambridge, Mass.: Acoustics Laboratory, Massachusetts Institute of Technology, 1952.

JELÍNEK, Petr. 2008. *Architektura a katalog informatických služeb.* Praha, 2008. 14 s. Referát. Univerzita Karlova v Praze, Filozofická fakulta, Ústav informačních studií a knihovnictví. Verze 1.0.

KAUFMANN, E.; BERNSTEIN, A. 2007. How useful are natural language interfaces to semantic web for casual end-user?. In *ISWC/ASWC 2007*. [s.l.] : [s.n.], 2007. s. 281-294.

KUHLTHAU, Carol Collier. 2004. *Seeking Meaning : A Process Approach to Library and Information Services*. 2nd edition. Westport : Libraries unlimited, 2004. 247 s. ISBN 1-59158-094-3.

KURLAND, Oren. 2009. Re-ranking search results using language models of query-specific clusters. *Information Retrieval* [online]. 2009, vol. 12, no. 4, s. 437-460. [cit. 2010-07-29] Dostupný také z WWW: < http://www.springerlink.com/content/h0756457qh093148/ >.

KWOK, Cody; ETZIONI, Oren; WELD, Daniel S. 2000. Scaling Question Answering to the Web. In *ACM Transactions on Information Systems.* Seattle : [s.n.], 2000. s. 1-22.

*Library of Congress Home***.** c2010. [online]. c2010 [cit. 2010-07-29]. Dostupné z WWW: <http://www.loc.gov/index.html>.

LIN, Jimmy, et al. 2003. The Role of Context in Question Answering Systems. In *CHI 2003* [online]. Lauderdale (Florida) : [s.n.], 2003. s. 1-2. [cit. 2010-07-29] Dostupné z WWW: < citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.8047 >.

LOISEL, A. 2009. *Modeling Human Interaction to Design a Human-Computer Dialog System* [online]. France, 2009. 6 s. Referát. LITIS Laboratory. [cit. 2010-07-29] Dostupné z WWW: < http://www.arxiv.com >.

MANKTELOW, Ken. 2000. ***Reasoning and Thinking*** . Hove (UK) : Psychology Press, 2000. 261 s.

MARTIN, Bill. 2005. Information society revisited : from vision to reality. *Journal of Information Science*. 2005, vol. 31, no. 1, s. 4-12.

McCLELLAND, James L. 2009. Is a Machine Realization of Truly Human-Like Intelligence Achievable?. ***Cognitive Computation***. 2009, no. 1, s. 17-21.

MENDEZ, Jovan D. R.; NAKAYAMA, Kenji. 2008. Adaptive Self-feeding Natural Language Generator Engine. In *IVA 2008 : LNAI 5208*. Heidelberg (Germany) : Springer Berlin, 2008. s. 533-534. ISBN 978-3-540-85482-1.

MERRILL, David, M. 2000. *The Instructional Use of Learning Objects : Online version* [online]. Ed. David Wiley. Utah : [s.n.], 2000 [cit. 2010-07-29]. Knowledge Objects and Mental-Models, s. 1-16. [cit. 2010-07-29] Dostupné z WWW: < http://reusability.org/read/ >.

MILLER, George A. 1995. *Classics in History of Psychology* [online]. Ed. Christopher D. Green. 1995. 1-15 s. The Magical Number Seven, Plus or Minus

Two: Capacity for Processing Information. [cit. 2010-07-29] Dostupné z WWW: < http://psychclassics.yorku.ca/Miller/ >.

MORVILLE, Peter; ROSENFELD, Louis. 2006. *Information architecture for World Wide Web*. 3rd ed. [s.l.] : O´Reilly, 2006. 504 s.

MOULIN, Bernard. 1998. A logical framework for modeling a discourse from point of view of the agents involved in it. In *ICCS´98*. M.L. Mugnier and M. Chein. [s.l.] : [s.n.], 1998. s. 359-366.

PAPÍK, Richard; PAPÍKOVÁ, Vendula. 2007. Informační chování ve věku online komunikace. In *INFOS 2007* [online]. [s.l.] : [s.n.], 2007. s. 1-11. [cit. 2010-07-29] Dostupné z WWW: < http://www.infolib.sk >.

PEREGRIN, Jaroslav. 2004. *Logika a logiky : Systém klasické výrokové logiky, jeho rozšíření a alternativy*. Praha : Academia, 2004. 205 s. ISBN 80-200-1187-0.

PILECKÁ, Věra. 2009. Informační věda v kontextu kognitivních věd. In *IKI 2009 : Informace, konkurenceschopnost, inovace 2009* [online]. [s.l.] : [s.n.], 2009. s. 1-25. [cit. 2010-07-29] Dostupné z WWW: < http://www.cisvts.cz/UserFiles/File/iki2009/Pilecka_ftxt.pdf >.

POLLACK, I.; PICKETT, J.M.; SUMBY, W.H. 1954. On the Identification of Speakers by Voice. ***Journal of Acoustical Society of America***. 1954, vol. 26, no. 3, s. 403-406 .

POPESCU, Ana-Maria, et al. 2004. Modern Natural Language Interfaces to Databases : Composing Statistical parsing with Semantic Tractability. In *Coling : 20th International Conference On Computational Linguistics.* Geneve (Switzerland) : [s.n.], 2004. s. 141-147.

POPESCU, Ana-Maria; ETZIONO, Oren; KAUTZ, Henry. 2003. Towards a Theory of Natural Language Interfaces to Databases. In *IUI´03*. Miami (Florida) : [s.n.], 2003. s. 1-9.

PORAT, Marc Uri . 1977. *Information economy : definition and measurement*. Washington : Washington Dep. of Commerce, 1977. 132 s.

PROUDFOOT, Diane. 2009. Meaning and mind : Wittgenstein´s relevance for the ´Does Language Shape Thought?´debate. *New Ideas in Psychology*. 2009, vol. 27, no. 1, s. 163-183.

RASMUSSEN, Edie M. 2003. Indexing and Retrieval for the Web. *Annual Review of Information Science and Technology*. 2003, vol. 37, no. 1, s. 91-124.

RYUTOV, T. A. 2007. Socio-cognitive Approach to Modeling Policies in Open Environments. In ***POLICY´07***. Bologna : [s.n.], 2007. s. 29-38.

SABOL, V., et al. 2002. Applications of a Lightweight, Web-Based Retrieval, Clustering, and Visualisation Framework. In *PAKM 2002*. Berlin : Springer, 2002. s. 359-368.

SAUSSURE, Ferdinand de . 2007. *Kurz obecné lingvistiky*. Praha : Academia, 2007. 487 s. ISBN 978-80-200-1568-6.

SIMON, Herbert A. 1981. Information-Processing Models of Cognition. *Journal of the American Society for Information Science* [online]. September 1981, vol. 30, no. 5, s. 364-377. [cit. 2010-07-29] Dostupný také z WWW: < http://onlinelibrary.wiley.com/doi/10.1002/asi.4630320517/abstract >.

SKLENÁK, Vilém, et al. 2001. *Data, informace, znalosti a Internet*. Praha : C. H. Beck, 2001. xvii, 507 s. ISBN 80-7179-409-0.

SKLENÁK, Vilém. c2009a. Robot. In *KTD : Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online databáze]. Praha : Národní knihovna České republiky, 2009- [cit. 2010-07-29]. Systém. Č.: 000000655. Dostupné z WWW: <http://sigma.nkp.cz/cze/ktd>.

SKLENÁK, Vilém. c2009b. Stop word. In *KTD : Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online databáze]. Praha : Národní knihovna České republiky, 2009- [cit. 2010-07-29]. Systém. Č.: 000000665. Dostupné z WWW: <http://sigma.nkp.cz/cze/ktd>.

SNOW, Rion ; JURAFSKY, Daniel ; NG, Andrew Y. 2005. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems* [online]. Cambridge : MIT Press, 2005. s. 1-8. [cit. 2010-07-29] Dostupné z WWW: < www.stanford.edu/~jurafsky/paper887.pdf >.

SOLOMON, Paul. 2002. Discovering Information in Context. *Annual Review of Information Science and Technology*. 2002, vol. 36, no. 1, s. 229-264.

SORICUT, R.; BRILL, E. 2006. Automatic question asnwering : Beyond the factoid. *Information Retrieval*. 2006, vol. 9, no. 1, s. 191-206.

SORICUT, Radu; BRILL, Eric. 2006. Automatic question answering using the web : Beyond the Factoid. *Information Retrieval*. 2006, vol. 9, no. 2, s. 191-206.

SPERLING, G. 1960. The information available in brief visual presentations. ***Psychological Monographs : General and Applied***. 1960, vol. 74, no. 11, s. 1-29.

SPINK, Amanda, et al. 2001. Searching the Web : The Public and Their Queries. *Journal of American Society for Information Science and Technology*. 2001, vol. 52, no. 3, s. 226-234.

STERNBERG, Robert, J. 2009. *Kognitivní psychologie*. Vyd. 2. Praha : Portál, 2009. 640 s. ISBN 978-80-7367-638-4.

STEYVERS, Mark; GRIFFITHS, Thomas L. 200?. *The Probabilistic Mind: Prospects from Rational Models of Cognition*. Oxford : Oxford University Press, 200?. Rational Analysis as a Link between Human Memory and Information Retrieval , s. 1-22.

SUSSWEIN, N.; RACINE, P.T. 2009. Wittgenstein and not-just-in-the-head cognition. *New Ideas in Psychology*. 2009, vol. 27, no. 1, s. 184-196.

TAYLOR, J.G. 2009. Cognitive Computation. *Cognitive Computation*. 2009, no. 1, s. 4-16.

THOMPSON, C.; PAZANDAK , P.; TENNANT, H.R. 2005. Talk to your semantic web. In *IEEE International Computing*. [s.l.] : [s.n.], 2005. s. 75-79.

TOMS, Elaine G. 2002. Information Interaction : Providing a Framework for Information Architecture. *Journal of the American Society for Information Science and Technology* [online]. 2002, vol. 53, no. 10, s. 855-862. [cit. 2010-07-29] Dostupný také z WWW: < https://www.unc.edu/~acrystal/110-117/toms.pdf >.

TREDDINICK, L. 2006. Web 2.0 and business : a pointer to the intranets of the future? . *Business Information Review* [online]. 2006, vol. 23, no. 4, s. 228-234. [cit. 2010-07-29] Dostupný také z WWW: < bir.sagepub.com/cgi/content/abstract/23/4/228 >.

***True Knowledge : the internet answer engine.* c2010** [online]. c2010 [cit. 2010-07-29]. Dostupné z WWW: <http://www.trueknowledge.com/>.

WILSON, T. D. 2006. Revising user studies and information needs. *Journal of Documentattion* [online]. 2006, vol. 62, no. 6, s. 680-684. [cit. 2010-07-29] Dostupný také z WWW: < http://www.emeraldinsight.com/0022-0418.htm >

Wolfram|Alpha : Computational Knowledge engine. c2010. [online]. c2010 [cit. 2010-07-29]. Dostupné z WWW: <http://www.wolframalpha.com/>.

***WordNet : Lexical Database for English.* c2010.** [online].Princeton University, c2010 [cit. 2010-07-29]. Dostupné z WWW: <http://wordnet.princeton.edu/ >.

YANG, Kiduk. 2005. Information retrieval on the Web. *Annual Review of Information Science and Technology* . 2005, vol. 39, no. 1, s. 33-80.

ZAÏANE, Osmar R. (1999) *Principles of Knowledge Discovery in Databases - Chapter 8: Data Clustering*. [online]. University of Alberta. Dostupný z WWW: <http://www.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter8/index.html>.

Evidence výpůjček / Evidence of lending

Statement:

I give permission to lend this thesis. The user confirms his signature; that this will work properly cite the list of the literature

Prague, 17 August 2010

Prohlášení:

Dávám svolení k půjčování této diplomové práce. Uživatel potvrzuje svým podpisem, že bude tuto práci řádně citovat v seznamu použité literatury.

V Praze, 17. 8. 2001

Barbora Poláková

Jméno	Katedra / Pracoviště	Datum	Podpis