Informational Linguistics

Informational Linguistics:

The New Communicational Reality

By

Alexander Barkovich

Cambridge
Scholars
Publishing

Informational Linguistics: The New Communicational Reality

By Alexander Barkovich

This book first published 2020

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-4424-9
ISBN (13): 978-1-5275-4424-6

CONTENTS

INTRODUCTION

The crystallization of the *informational linguistics* format is predetermined by the undefined knowledge content of natural sciences in the professional competence of philologists. Communication has become computer-mediated, and the social and cultural development of mankind depends on network technology but everything seems to remain the same in the world of humanities as in the Sleepy Kingdom. It looks like linguistics is redundant in times of technogenic and dynamic progress and is just an anachronism from the scientific romanticism era. Nevertheless, there is a paradox: No science is as innovative and useful in the modern world as linguistics is. With one provision: this linguistics must be informational.

Language functioning in the modern social and technological circumstances requires not only the study of speech samples in separate chronological or spatial frames but also the wide *systematization* of speech practice. The modern pace of scientific and technological development contributes to significant speech expansion. This expansion is supported by the computer, the revolutionary character of which is compatible with the wheel and fire. Language is changing due to the computerization of human interaction. No living language has remained uninvolved in the developments in the social and cultural spheres: Millions of texts are stored on servers operating on the Internet.

Today the questions of language system positioning in a multifaceted symbiosis of knowledge are more important ever. Interdependence and the syncretism of human cognitive and creative activity – in the face of rapid technological process – create unlimited demand for knowledge interpretation in a universal format. This format is *linguistics*. It is linguistics that has always been an important source and reserve for scientific development.

> We observe at the same time that these new methods of linguistics are taken as examples and even as models for other sciences, that the problems of language are today of interest in very diverse and increasingly numerous fields of specialization, and that there is a trend in the research done in the social sciences toward working with the same mind that inspires linguistics. [30, p. 17]

As practice shows, each individual computer system, the *World Wide Web* as well as the *Internet,* could not exist and develop without *linguistics*. Not only is the communication saturation of the social networks relevant here but also their dependence on *language* as the main and only instrumentality of human interaction. Language as a phenomenon is diverse – from gesture to poetry – but one way or another it is a code system, dependent on the means of its "maintenance" – linguistics. It is not surprising that all of us are competent in linguistics at least to some extent.

With that, the linguistics of modern communication is more hidden due its total clichédness and stereotypy; it is equipped with mechanisms like the radio, TV and computer. Today, the meta-language practice of programming is another typical kind of new linguistics. Even the measurement of such modern communication as "virtual reality", which seems modern and original, is apparently conditioned by linguistic well-formedness. Today the main peculiarity of communication is probably its *informatization*. Step by step *information* has become the main tool of communication. Of course, it has always existed but today we cannot live without it, like we cannot live without modern clothes.

Meanwhile, with the advent and spread of a fundamentally new format of language fact functioning – in the communicational networks environment – linguistics now includes material of previously unattainable quality and quantity. With speech-processing applications such as *corpus* and *Internet search*, the representativeness of *language material* has become provisional: it is possible to explore either several language units or millions of them in the same way. Step by step, the issues of card file aggregation and their adjustment (computer programs cope with this task successfully and instantly) were replaced by issues surrounding the formation of research strategy and tactics. Today the essence of speech phenomena and processes is in focus, as well as their correlation with the world. The existing views and the conceptual representation of such understanding rely mainly on *information*.

The hyper-phenomenon of modern speech practice, the Internet, is directly related to linguistic theory and practice: it is a communication network. The Internet has become a real catalyst for the active development of all current scientific paradigms. The *Internet* is a prominent linguistic object: It is a non-trivial example of semantic expansion that takes part in the wide replacement of related language units or generalized categories. On top of that, the linguistics of the Internet, or *Internet linguistics*, is quite a self-sufficient discipline. This linguistics is not of the classic type. First of all, this Internet-driven linguistics is *informationally* determined. Today it is obvious that narrowly focused linguistic paradigms are not able to catch modern speech dynamics, their manifestations and sense. Now is the time for the new *methodology* of linguistics that is based on cognitively verified knowledge. It unifies *a wide range* of humanities competence, is supported by computer instrumentality, and is complex and multilevel.

Current research is often carried out on the boundaries of different disciplines or branches of disciplines. There are different explanations for this situation: Firstly, science is a *synthetic* phenomenon artificially divided into narrow segments; and secondly, the results of a search activity are often simultaneously useful for a number of disciplines. *Inter*- and *multidisciplinary* methodologies have undeniable perspectives in the context of modern technology and humanity trends. They are already being successfully implemented in such fields as cybernetics, artificial intelligence and web science, which are not self-sufficient and homogeneous. In the same way, research directions gradually merge into a huge problem domain that is united by the specifics of *informational linguistics*.

Contemporary information has an idiosyncratic identity, with scientific substantiation of related issues, meta-structural ties, and their interdependence. Two main dimensions are distinctive in the syncretic domain of informational linguistics: communicational and discursive. The computer mediation of information has become a sign of the times, allowing the identification of particular kinds of *paradigmatics* – sets of aspects and instrumentality. In the modern understanding they are known as *computer-mediated communication*, or *CMC*, and *computer-mediated discourse*, or *CMD*.

There are many unsolved problems in linguistics as well as in communication science. There are many blank spots that were not touched upon and or even mentioned but could be identified in the syncretic *informational* problem domain. N. Baron, D. Crystal and S. Herring are among the pioneers here, forming the frames of a new computerized disciplinarity. The problematics of contemporary speech activity – *discourse* in particular – interest many researchers, among whom are T. van Dijk, D. Schiffrin, W. Chafe, R. Wodak and N. Fairclough, but still the conceptual description of the informational specificity of modern speech, computerized speech especially, has sufficient unengaged potential.

Contemporary linguistics is better equipped with statistical instrumentality and is characterized by an increased methodological mobility and pliability. Interesting and promising findings constantly appear on the border of humanities and technical knowledge. The objects of study and modeling in artificial intelligence, the World Wide Web, the Internet and other fields are often difficult to separate according to certain disciplines and sub-disciplines. Contemporary linguists allow the identification of its *crossdisciplinary*, *multidisciplinary* or *interdisciplinary* methodological character. With that, most objects of study develop in the trend of synthetic scientific exploration but due to contemporary linguistics' communicational attributability, they are “linguistic-centered”, comprehended within the *linguistic* paradigm.

Linguistics has some exclusivity in science: it directly or indirectly covers all human activities, including technical and computer fields. Thus, software programs, mathematical theorems and many other scientific phenomena are expressed through the instrumentality of linguistic (meta-language) means. Moreover, all modern technological progress is clearly mediated via informational and communicational conceptualizations, which first have a linguistic nature. Obviously, many of the technological advances or setbacks are due to sufficient or insufficient meta-language correctness and the degree of *linguistic competence* of “operators”. It is linguistic or metalingustic support that is in strong demand in the innovation processes of the symbiosis of science and production.

On the other hand, such support in any activity requires consistent and uncontroversial *scientific pragmatics*, systematization and representation of data. This kind of linguistics must be different: first as correct as possible and only then beautiful. This era of computers and network technologies has changed life and, of course, linguistics irrevocably. The long-term perspectivity of modern linguistics is provided with syncretic theoretical-practical activity without artificial, superfluous limitations.

But there are natural peculiarities as well. One of them is the *polycode* character of modern speech caused by the wide use of multilingual material, for example, the Latin inclusions in English. But English on its own is a much-in-demand source of such inclusions for other codes. Most languages are involved and exemplified with related speech practice. Cyrillic languages (such as Russian, Belarusian, Ukrainian) are interesting linguistic objects in this regard: they allow speech specificity without the interference of many graphically similar languages (particularly Germanic or Latin ones), which are to a large extent related. By the way, it is the Russian language that ranks second (after English) in prevalence on the Internet [209].

In addition, there is “unusual” speech: programs function along with traditional natural-language texts. They are being reluctantly considered as speech practice, though the related activity is based on languages – artificial and formal but languages nevertheless. Moreover, programs are used for processing and meta-describing natural-language speech, causing *software* dependence. Speech practice with software support preserves the great authenticity and spontaneity that add special value to the linguistic meta-descriptions. The computerized or *digital* format of supported texts is its inevitable requirement; their markup (or annotation, tagging) is optional. Mastering such quality characteristics as accessibility, pre-elaboration (or data pre-processing) and adequacy of text processing techniques to the task matters; and others are at the forefront.

Additionally, the qualitative specificity of the problem domain is complicated by its *meta-textual* – *hyper-*, *inter-*, etc. – structure. The current speech activity is *global* and *total*: a word can have unexpected and deep effects. Any new meta-description has to take the previous one into account or should affect it in turn, creating a *blockchain* quality of communication. A distinctive feature of the relevant problem domain is the linguistic capability to simultaneously involve the widest possible range of illustrative material via computer technology. With that, printed texts keep their importance: dictionaries, for example, which are still under-represented in computer ("electronic") format. One way or another, it is quite simple technically to change the printed format into a computerized one and vice versa.

In the new communication conditions, significant linguistic relevance is provided by strengthening the *above-personal* identity of speech practice "enforced" by text corpora, Internet discourse, digital libraries, and online dictionaries. Most of this speech practice is already separated from its authors. This character is manifested in both the organized format of *institutionality* and the free format of *noo-spherity*. For example, the popular linguistic instrument of text corpora is created for most languages. These language dimensions are *astronomical* in quantitative measurement.

The representative stability of contemporary language data allows its purposeful study and the creation of widespread meta-descriptions and *generalizations*. Such theoretical generalizations lend credibility, which is desperately needed for the entire field of linguistics as well as for many neighboring spheres. In recent years new methodological techniques have become available, providing more knowledge for the *analysis*. With that, the ambiguity of the current scientific sphere is badly in need of *synthesis*: knowledge must be gathered as an organized, described and accessible object.

Such options of knowledge mastery are possible with the universal communicational matter, or "currency" – *information*. Information creates the phenomenological basis for the formation and development of new cognitive "religion" or ideology – the informational one. This ideology, in turn, is indispensable with the linguistic paradigm: it is *linguoinformational*. But the objective side of this aspect still requires systematization and methodological definition, and practice goes faster here than theory. It is impossible to correct tactics, to comply in a comprehensive manner with a holistic and integrated research strategy without such coherence. Generally, it allows the preservation of the conceptual consistency of the whole related activity.

Informational generalizations are really in demand. It would probably be advisable to define *information* as an essentially new substance of communication. With that, at first glance it is no different from the well-known *semantics* – the "prima materia" of speech interaction. But the content in the guise of information has a certain aura of newness and authenticity, attracting the imagination of explorers; its mysterious appeal is very useful and attracts neophytes. Of course, the inexorable logic of the material world suggests that the shiny informational shell of computer-mediated communication is just the cocoon of an enchanting butterfly of sense that always escapes "entomologists", both amateurs and professionals. But the magic of the art of word control has hypnotic power.

No one has ever seen information or language but just as the ephemeral nature of language does not interfere with the existence of linguistics, the ephemeral nature of information is not an excuse to ignore *informational linguistics*. Information is an abstraction, like everything else in the cerebral world. With that, it has already gained the critical mass of conceptualization, and in turn promotes distinct identity acquisition for ***informational linguistics*** – a fully-fledged discipline dedicated to the investigation of the specifics of communication contents [23].

The modern linguistics problem domain is huge and touches upon other scientific spheres. Nowadays linguistics is doomed to be applied linguistics. It manifests in the structuring of such syncretic directions as psycholinguistics, sociolinguistics, forensic linguistics and dozens of others that "appear" annually. However, the novelty of such a process is rather superficial: Language, speech, communication, and discourse are universals that have always existed and have always interfered with the human perception of the world; as such the *informational actualization* of linguistics is an inevitable and consistent stage of knowledge gain.

Meanwhile, experience has proven that the road to informational linguistics is difficult and this specialization is quite new for universities, but at the same time other "universities", such as life, for example, attract many newcomers to the field of informational linguistics. These specialists differ in qualifications but are equal in their passion and readiness to change the world and the future. So, announcing the *visualization* of informational linguistics is not essential – it could be considered as a new generation of informatics or a new branch of linguistics. But informatics – like *information technologies* themselves – has always been considered an abstract *metaphor* for communicational activity. In contrast, informational linguistics is more real and actually a whole unit. There is no doubt that the informational dimension of communication is relevant first of all from the linguistic point of view here. Its essence is important for every field of human interaction.

It is obvious that only by means of linguistic instrumentality can hindrances be removed and the way to genuine artificial intelligence opened. Everybody knows what it is even though it does not exist. With that, it would be a mistake to deny artificial intelligence or make it a fetish. The worst mistake is the attempt to assimilate human mentality with the "artificial" communicational standard. Other than the fact that it does not exist, it will be *different* and able to simplify human life to a mechanical level. Nobody would agree, for example, to change the ability to see colors to just seeing black and white but some natural-artificial *intellectual interfaces* look quite real. Common environment or communicational reality for natural and artificial intelligences could be informational, for example; at least no other alternatives are visible.

In the context of these circumstances, informational linguistics occupies a *special place*: it is on the cutting edge of scientific and technological progress and integral to our daily lives – by switching on a computer and connecting to the Internet, we are studying it anyway. The only issue is to study it efficiently with our eyes open.

CHAPTER 1

INFORMATION AS A LINGUISTIC PHENOMENON

1.1 Conceptualization of information

Information is one of the key concepts of the communicational sphere nowadays but the *informational* specificity of modern communication is challenging: Information is positioned in many linguistic descriptions as a very contradictory substance, particularly in *meta-language* understanding. In fact, *meta-language* is a secondary semiotic system for any language description in a wide sense. But, naturally, *meta-language* in the computer-mediated environment is hardly conditioned by the discreteness of formal languages. Of course, any relationship in CMC must be linear and clear. A meta-language is traditionally based on *vocabulary*, or lexis, but, simultaneously, speech has some additional super significance; for instance, contextual. It must be identified for CMC otherwise the computer cannot use it.

After the lexical level, syntax is the next and last "visible" language level. Naturally, in CMC conditions it is mainly in syntax that all the deficit of speech *significance* is looked for. Really, much of speech semantics depends on the *syntagmatics* of the software "texts". But it is right here that the misunderstanding of real meta-language functionality begins. First of all, *syntax* is linguistically sophisticated on its own, even if such relationships do not seem relevant for computing processing. Then, it cannot substitute all the semantics. According to such misunderstanding, *information* is only a concept, and there is no intention to imbue it with any genuine sense. It is convenient and does not prevent the systemic associating all unresolved problems of speech meta-descriptions – especially CMC-related – with *information* as a semantic concept [21]. But information is not lexical or syntactical, or anything else artificially framed – it is the essence of all the language involved in communication.

In many aspects it is reasonable to ***interpret*** information as a phenomenon. As a notion it is integrated with *communication* and closely connected with it in a linguistic context. First of all, communication reflects the dynamism of language functioning, providing its multileveled semanticity, which is absolutely inherent to information. Communication paradigmatics allows the real meta-language operating of information, which is what makes related research practice linguistic-conditioned. The functionality of information in the field of CMC has another benefit: its identity is supported by powerful statistical possibilities.

Then, the notional interpreting of information can be consistently developed at the next stage – the process of its conceptual ***representing***. CMC information does not match only the traditional understanding of this concept; it also requires additional meaningful clarification and interpretation. Of course, the communicational interpretation of information is not sufficient for the speech practice aspect and needs to be supported with a widened categorial systematization. This is an important aim – the describing functionality of *information* – as a meta-language phenomenon – is largely due to its conceptual representation. Thus, it is advisable to construct objective descriptions of creation, transition and usage of information on the basis of comprehensive and universal *methodology*. The consistent methodology allows reliance on the proven argumentation – keeping the intralinguistic logics. It is *a discourse* that has such particular authenticity. After all, discourse as a paradigm allows the comprehensive investigation of the extra-language specifics of information functioning. A discursive prism provides interdisciplinary value for any speech ***modeling***, making it compatible with all scientific knowledge as well as its applied specifics.

The form and content of the lexeme *information* in the modern sense did not appear as an instant and unprepared linguistic innovation. It has a long and winding path of language development – *semantization*. The etymology of *information* is associated with the Latin word *forma* – 'form, contour, figure, shape; appearance, looks; a fine form, beauty; an outline, a model, pattern, design; sort, kind condition'. There are reasonable grounds for recognition of the Greek lexeme *μορφή* (*morphḗ*) – 'form, beauty, outward appearance' – as a prototype unit of the Latin *forma*. Later the derivant *forma* was actively involved in word-formation processes in Latin: *forma* created the verb *formare* ('to form, shape'). Joining the prefix *in-* to the *formare* made *informare* ('to shape, give form to, delineate', "figuratively" 'train, instruct, educate') one of its derivations. Afterwards in Old French *informationem* – nominative *information* ('outline, concept, idea') – a noun of action from the past participle stem of *informare* ('to train, instruct, educate; shape, give form to') – was transformed into the lexeme *informacion* (*enformacion*), meaning 'advice, instruction'. In the 14th century, this lexeme gained the meaning 'act of informing, communication of news', and became typical for the English calque *information*. The meaning of ‘knowledge communicated concerning a particular topic’ has been associated with the lexeme *information* since the 15th century [156]. The terminological word-group *information technology*, or *IT*, which was introduced in 1958 (coined in *Harvard Business Review*), is perhaps the most popular analytical development of *information*.

The lexeme *information* has become internationalized in the scope of the dynamic development of IT and the CMC-sphere. There is no problem in customizing the same lexeme exponent in most modern languages – it functions in the majority of languages in a form very similar to the original. Its meaning is quite standard, too. The semantics of *information* has undergone many modifications in the process of speech practice, but one cannot assume with certainty that the semantization of this notion is complete and the final significance already formed.

Until the middle of the 20th century the meaning of *information* was very common and vague: ‘knowledge’, ‘facts’. With the appearance of the CMC environment and the formation of the research area of the problem domain, its content plane acquired many shades: *information* became comparable with such fundamental scientific concepts as *matter* and *energy*. It has become an extraordinary multi-pronged concept – leaving behind, for example, the banal *data* – and continues to evolve wider and deeper [64]. ***Information*** is the meaningful specifics of communication, the conceptual-speech quintessence of language units in context combination.

N. Wiener, the father of Cybernetics, explains the essence of information through its functionality: “Information is a name for the content of what is exchanged with the outer world as we adjust to it, and make our adjustment felt upon it” [216, p. 17].

Without exhausting all the semantics of the modern understanding of information, this definition illustrates the “technological” approach to the interpretation of communicational content. The main drawback of this approach is transferring the significance from the human relations sphere to the supposedly independent “external” world of human communication. It presumes that, according to the proposed logic, the laws acting in technical and human communications are different. The next logical step in this direction could be admitting the need to adapt human communication to artificial standards. This step is often taken because it seems quite simple. There have been many attempts to simplify human interaction down to a computer-acceptable level in practice and it has so far proven to be impossible: heuristics is too deep for a too-shallow algorithm. In practice, nobody can fully align technical and computer-mediated communications. Thus, there are indicative limitations for the significance of the lexeme *information* in the world of algorithms.

> The word information, in this theory [statistical], is used in a special sense that must not be confused with its ordinary usage. In particular, information must not be confused with meaning.
>
> In fact, two messages, one of which is heavily loaded with meaning and the other of which is pure nonsense, can be exactly equivalent, from the present viewpoint, as regards information. It is this, undoubtedly, that Shannon means when he says that "the semantic aspects of communication are irrelevant to the engineering aspects". But this does not mean that the engineering aspects are necessarily irrelevant to the semantic aspects.
>
> To be sure, this word information in communication theory relates not so much to what you do say, as to what you could say. [212, p.8]

This explanation helps us cope with a specific task but decreases the chance to solve the general problem as such.

Meanwhile, such explanations are too simple and narrow. Facts or data are clearly useless for the linguistic (and realistic) representation of *information*. Moreover, they don’t work in the CMC context. Here, alongside well-known and obvious informational characteristics such as ‘obtained from investigation’, ‘representing data’, ‘which justifies change in a construct’ [63], *information* has gained new attributive computer-mediated characteristics such as ‘circulating in a network environment’, ‘contained in electronic format’, ‘available via the Internet’, etc. Today the modern conceptualization of *information* is inextricably linked with the dominating type of interaction, being actualized by it: information “lives” in the processes of creating, storing and transmitting messages in the computer-mediated environment.

The complex of essential development in the aspect of the modern communication includes, first of all, the issues of structuring, preserving and processing of information. CMC not only objectively contains information but makes it more visible. The CMC environment is characterized by the availability of relevant empirical material for any large-scale research, providing it with effective technological support. The informational relevance of the *large-scale* empirical material is predetermined by its quality of “quintessence,” which may be prominent only in the background of a fairly large quantity of material. With that, enormous quantities of aggregated speech provide many aspects of linguistics with new quality, allowing speech (and language) to be considered under the “microscope” of its *significance*. Such a “microscope” of the significance is as useful as the “telescope” of the functionality mentioned by T. McEnery and A. Hardie [144].

The communication sphere includes a huge amount of information that is constantly functioning and being updated. Though related mechanisms are still unknown, this does not prevent the human mind from processing successfully. Of course, human mechanisms of information processing have little to do with duplication; nevertheless, they work correctly and are constantly improving. Today we continue the ancient process of complicated multi-channel communication, started by the first humans. The quality of “natural information”, inherent to human intellect, and “artificial information”, inherent to computer programs, looks very different. It is time to rationally separate their terminological registrations and rigorous metalinguistic descriptions. So, if ***natural information*** reflects the contextual potential of the intuitive kind for heuristic interpretation in the natural-language environment, ***artificial information*** is the algorithmically correct textual add-on for processing speech in the discrete environment of formal languages.

The effectiveness of computer facilities is incomparably superior to the statistical and algorithmic abilities of people. In this connection, information functioning in CMC is naturally conditioned by its meta-language representation. The possibilities for describing the “essential” aspects of modern communicational activitities are irreversibly mediated by the computer technologies sphere. No researcher would change the computer keyboard for a typewriter or a fountain pen.

1.2 Specificity of CMC-information

The statistical model of communication, which creates the digital world and *artificial information*, is based on the categorical necessity of choosing between two variants of abstract *binary opposition*. At the *bit* level this is known as “0” or “1” (“yes” or “no”). With all the limitations of this model, it allows the description of the communication processes in a discrete coordinate system in the context of unified, logically justified equivalents. This approach is the background of general theory of control and connexity based on the statistical measurement of communication, or *Cybernetics*.

Nobel laureate Dennis Gabor (or Dénes Gábor) described the *essence* of the *statistical model of communication* in 1952.

> Once we have a vocabulary, communication becomes a process of selection. A selection can always be carried out by simple binary selections, by a series of yeses or noes. For instance, if we want a letter in the 32-letter alphabet, we first answer the question "is it or is it not in the upper half?" By five such questions and answers we have fixed a letter. Writing *1* for a "yes" and *0* for a "no", the letter can be expressed by a symbol such as *01001*, where the first digit is the answer to the first question, and so on. This

> symbol also expresses the order number of the letter (in this example, the number *9*) in a binary system. [82, p. 1]

It is this model that provides the functionality of information in CMC. Such instrumentality has macro levels, for example, *programs*, as well as micro levels, presented by a specific sub-model called a *bit*.

Naturally, this instrumentality allows the formalization of only a superficial shell of communication, for example, graphics or acoustics, and needs to be elaborated for further adjustment. The *ontological* problem of improving the *statistical model of communication* is the impossibility of objective representation of speech practice by algorithmic procedures of computer mediation. It is too extensive and needs a special generalization of material for effective mastering – information. Moreover, speech practices, including computer-mediated ones, are multidimensional and variable. The correctness needed at the level of abstract modeling of communication mechanisms is scarcely supported by the adequate representation of such involved semiotic systems as natural languages. Nevertheless, it is very "convenient" from a technical point of view: the issues of baseload meta-description, for example, *dictionary*, and its replenishing, its mastering with new participants of communication, for example, children, are taken out of brackets. In such a mode human senses are called "chaotic"; for no reason they are accused of *interfering* with the describing and understanding of information [82, p. 1].

It is *au contraire* in linguistic practice. Of course, the effectiveness of computer tools is high but the capabilities of CMC only *complement* traditional methods of communication interpretation. The metalinguistic structuring of "thinking" in CMC does not significantly differ from pre-computer speech practice. However, the computer presentation of semantics is based on the formal logic of special tables of commands, *programs*, which is not directly "compatible" with the intuitive mentality ("heuristics") of a human being. *People's mental* activity is significantly different from computer information processing, and the differences are clearly visible through the prism of speech functionality.

> First, and perhaps most important, is a confusion about the notion of "information processing". Many people in cognitive science believe that the human brain with its mind does something called "information processing", and, analogously, the computer with its program does information processing; but fires and rainstorms, on the other hand, don't do information processing at all. Thus, though the computer can simulate the formal features of any process whatever, it stands in a special relation to the mind and brain because, when the computer is properly programmed, ideally with the same program as the brain, the information

> processing is identical in the two cases, and this information processing is really the essence of the mental.
>
> But the trouble with this argument is that it rests on an ambiguity in the notion of "information". In the sense in which people "process information" when they reflect, say, on problems in arithmetic or when they read and answer questions about stories, the programmed computer does not do "information processing". Rather, what it does is manipulate formal symbols. The fact that the programmer and the interpreter of the computer output use the symbols to stand for objects in the world is totally beyond the scope of the computer. The computer, to repeat, has a syntax but no semantics. Thus if you type into the computer "2 plus 2 equals?" it will type out "4". But it has no idea that ‘4’ means 4, or that it means anything at all. [180, p. 202]

In the process of human perception of *natural information*, for example, in reading, specific mental mechanisms are involved. *Reading* in the traditional sense of the word is a unique activity of the brain that is effective not due to the rapid recognition of a number of images but because it is characterized by the slow mastery (and preservation throughout life) of the essence of things and concepts, or information. The *cognitive* specificity of mentality is due to the information storage process inherent in people.

> the stored information of the mind lies on many levels of accessibility and is much richer and more varied than that which is accessible by direct unaided introspection. [217, p. 149]

Unaided introspection in cybernetic understanding is the work of our sense organs. Artificial information, unavailable to us for direct introspection – for example, derived from *Internet surfing* – must obviously acquire a “natural” quality in order to be available for processing by the mentality. But only ascertaining the objective difference between the learning of information by the human brain and the processing of data by artificial intelligence programs is not enough. We should *combine* them.

> As we have already seen, it is not the empty physical structure of the computing machine that corresponds to the brain – to the adult brain, at least – but the combination of this structure with the instructions given it at the beginning of a chain of operations and with all the additional information stored and gained from outside in the course of this chain. This means that not only must the numerical data be inserted at the beginning, but also all the rules for combining them, in the form of instructions covering every situation which may arise in the course of the computation. [217, p. 146]

But it is here – on the threshold of the multi-channel intuitive logic of the human mentality – that computer programs have stopped. Forever or not, it is a matter of the continued existence of CMC: it must develop or it will be substituted by another invention. Still, an important scientific issue is the search for a unified standard platform for communicational interaction. Yet the lack of evidence of such consensus allowed Sergey Brin, one of the founders of *Google*, to note the fundamental imperfection of processing the "informational" content of communication both by the human intellect and artificial intelligence programs (in an interview with *Newsweek magazine*), "Certainly if you had all the world's information directly attached to your brain, or an artificial brain that was smarter than your brain, you'd be better off. Between that and today, there's plenty of space to cover" [129].

In this context *CMC-information* is in demand as some kind of a "cipher" equivalent. Scientists began measuring information – as an abstract computer equivalent of semantics – in the middle of the 20th century: the term *bit* was introduced in 1948 and *byte* in 1956. According to the *statistical model of communication* – which formed the basis of the modern approach to CMC structuring – the informational unit, or the unit of information, must correspond to one choice between alternative variants. A ***bit*** is such a unit. A ***byte***, traditionally a group of eight bits ("octet"), can have, respectively, 256 (2^8) variants, or "meanings". For example, according to all the data in the world as of 2011, "information" amounts to about *2.56* zettabytes (the prefix *zetta* means multiplication by 10^{21}). As of May 2015, the total number of digitized data in the world exceeded *6.5* zettabytes, and by the end of 2015 it exceeded *8* zettabytes. It is predicted that by the end of the 21st century, the information in the world will amount to *4.22* jottabytes (the prefix *jotta* means multiplication by 10^{24}) [220]. However, the development of "hyperinformation" will obviously be ahead of any forecasted rate. It is quite possible that by the end of the 21st century, the amount of informational "mass" will be much larger and it will be necessary to introduce new units of measurement. At the same time, it is more correct to talk about the mechanical accumulation of *data* in the aspect of CMC.

Moreover, what is not known to humans (and not reflected or envisaged in the meta-language generalizations) cannot be used by computers. Therefore, engineers, when duplicating human-like mentality consciously or subconsciously, associate it with the mentality and speech practice of human beings. Finally, the effectiveness of CMC functioning undoubtedly depends on the extent to which such associations are supported by the purposeful and conscious use of the meta-language apparatus. This is why a special separated world of computer technology and *artificial intelligence* has not been created yet, and it is not known whether it ever will.

Today, one way or another, scientific activity in CMC is associated with the mastering of *data*. At the same time, this practice has developed due to the demand for suggested verified outcome – ***knowledge***. But true informational results are hard to achieve while information is an abstraction and it is impossible to lose or find it materially. We often treat "information" as a typical metaphor of "data". Of course, the necessity *to mine* some data and its *retrieving* are quite relevant but successful *data mining* is of little use for the mining of *information*, which could be represented by some signs but, finally, has to be represented in human mentality – as something abstract. The computer has no capacity for such transformation.

With that, computer capacities are massively enlarged with communicational networks. This has been proven by the *World Wide Web*, for example.

> The Web is a principled architecture of standards, languages and formalisms that provides a platform for many heterogeneous applications. [35, p. 16]

But again nets only support the accumulation or aggregation of data. One way or another, such aggregation is now limited and could be fatal in future. Such ambiguity is quite understandable and predetermines active efforts to look for the way out of this situation. One of the supposed directions here is the idea of a new generation of communicational network – the *Semantic Web*.

> However, the Semantic Web, a vision of extending and adding value to the Web, is intended to exploit the possibilities of logical assertion over linked relational data to allow the automation of much information processing. [35, p. 5]

The ideology of creating a Semantic Web is based on achieving a qualitatively new level of data: generalized knowledge, or information. Its founders-architects argue

> The Semantic Web (SW) is an attempt to extend the potency of the Web with an analogous extension of people's behaviour. The SW tries to get people to make their data available to others, and to add links to make them accessible by link following. So the vision of the SW is as an extension of Web principles from documents to data. [35, p. 18]

Why hasn't it been created yet? Because there must be a movement not "from documents to data" (already done 70 years ago) but "from data to knowledge" – via information.

Thus, the next decisive improvement of CMC depends on the acquisition of information mastering. The informational equipment of CMC is extremely important for the development of modern communication. The main "drawbacks" of CMC are still due to the "shallow" semantics of computerized communication. The interpretation of CMC only superficially correlates with the wide and multilevel meaning of any sign or signal. One way or another, in the imagination of communicants – and in the programs of their servants, computers – these meanings are self-sufficient. But this illusion is not acceptable for science. Moreover, remaining with the connectivity reflected in communication reality, any meaning should be abstract. How to attain this? Through the "linguistication" of communication and "informationalization" of communicational reality.

1.3 Limitations of informationalization

Of course, the invention of the computer, the World Wide Web, the Internet, and many of their attributes lifted the state of modern science to fantastic heights compared to the mid-20th century. It seemed that at some point the science changed fundamentally again. A lot of superfluous theoretical knowledge was rejected as too old in the name of quick super-goal achievement – the creation of artificial intelligence. Surprisingly, some new practical data worked successfully even without steady theoretical support. Up to a point. Today the last inventions rather create new problems than solve old ones. It is obvious that the time of data is over: theoretical knowledge is in demand again.

By the way, the World Wide Web was created long ago enough for today's pace of evolution – about 30 years. Nothing principally new has appeared in the communicational sphere since then. Herewith practice testifies that it is not possible to maintain the same dynamic pace of technological development, determined by the needs of communicational sphere. Moreover, some "old" important projects in the CMC-sphere have still not been realized; for example, sustainable *machine translation* was promised in the 50s but has still not been achieved. And it looks as if scientists will work on it for at least 50 years more.

Permanent adjustments of the communicational sphere are quite understandable. Human development is spiral, and now it looks as if the times of René Descartes and Gottfried Wilhelm von Leibniz are coming again: theory is in demand today. This stage is more sophisticated than before: it is not monodisciplinary but rather symbiotic and interdisciplinary, syncretic and multidisciplinary. In other words, today's communication should become linguistics-driven, and linguistics should focus on communication. The potential for a formal kind of quasi-linguistics, which until recently felt right at home in informatics, has been exhausted. Now a simple question is on the agenda: how is it possible to integrate semantics, discourse, and other fundamental universals into communication? At the same time, there is a serious burden: nobody wants to annul anything that has already been done. Such a bridge could be constructed on informational grounds.

In the theory of language, the role of the information hyperonym *semantics* is well known. The attempts to formalize these important relations were being made permanently. It seems that there is a visible basis.

> Briefly, such is the method of distribution: it consists in defining each element by the ensemble of the environments in which it may occur and by means of a double relationship-the relationship of the element with the other elements simultaneously present in the same portion of the utterance (the syntagmatic relationship) and the relation of the element with the other elements which are mutually substitutable (the paradigmatic relationship). [30, p. 102]

Really, it looks quite understandable: there are just two dimensions in the language coordinate system – "vertical" *paradigmatics* and "horizontal" *syntagmatics*. If we describe it consistently, we can get the grammatical picture of the language. It will indeed be the grammar. Of course, this recipe was proven in the formal world of artificial communication but computerized languages don't work in this mode. The matter is that this is not the whole of grammar. There is something else again – *meaning*, e.g. grammatical meaning. This additional "magic" component is semantic, or informational – as its meta-substitution. But this is inevitability still often considered as too theoretical, and is overlooked when the mechanisms of communication are looked for. Information, the inexhaustible spirit of language, determines the variability of communication in practice. Particularly, there is an intuitive *mechanism* of language that determines its life – communication. With that, it is technically impossible to describe it using superficial algorithms today.

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual

> entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design. [183, p. 31]

What we still have is the ineffectual substituting of the multilevel logic of human informational processing with its simplified shadow – dictionary representation. It works precisely in the artificial reality of databases but does not work in interaction with human participation. Computers can operate with different kinds of meaning as long as they are frozen and separated from practice ones. This conclusion, made in the middle of the 20th century, is relevant still. The problem will remain unsolved until meaning is considered as additional to "pure" communication.

Another issue is the criteria of communication success. They are often subjective. But there will hardly be a reason to consider the project as "successful" when it is not finished and its outcome is close to 50/50. At the same time, it is possible in such "applied" linguistics as informatics. For example, the issues of automatic recognition of anaphoric relations were "successfully" solved in this way: "The overall success rate calculated for the 422 pronouns found in the texts was 56.9% for Mitkov's method, 49.72% for Cogniac and 61.6 % for Kennedy and Boguraev's method" [130, p. 24].

One way or another, it is interesting that the result can be even less than 50% – "49.72%." Anyway, the indicator of anaphora recognition in English by computer software is not very high: the "overall" success rate hardly reaches 77% [130, p. 37].

For Norwegian, language anaphora recognition showed similar figures – from 68.92% to 74.6%.

> Nevertheless, the accuracy of 74.60% obtained by the machine learner on the development corpus is significantly better than the 68.92% accuracy achieved by my ARN reimplementation, and it is also much better than the 70.2% obtained by Holen's original system on the older version of data drawn from the same corpus. [153, p. 252]

According to this report, different options were used on the basis of text corpus, including *machine learning*. However, the computer "pupil" acted even worse than the human "teacher". On the base of the Czech National Corpus, the results were totally comparable: from 61.5% to 69.5% [152].

Thus, the perspectives for getting high quality are not clear and a *perfect* result in the formalization of discourse system properties is still far from being attained. In the semantic aspect, the automatized mastering of speech does not provide acceptable solutions. Here the programs failed to even reach the 5% "noise" threshold, which is usual in the manual processing of speech practice. After all, the percentage of accuracy from 60% to 80% is generally similar to the accuracy ratio of 50% in the *Probability Theory*. In practice such "results" of computer programs are hardly satisfying for programmers or their customers. By the way, this problem is quite serious for linguistics and is still just fractionally described but for CMD it turns out to be "solved" – without the *theory of language* again. This desire is quite understandable: anaphoric relations are sufficient for speech modeling. But with that they are a sophisticated task for formalized meta-description as the relations are not linear or determined by deep semantics.

Simultaneously, anaphoric relations are not the only problem of CMD, which is growing and developing all the time. This dynamics further complicates mastering speech: "Above all, though the World Wide Web in many studies can be considered as a static space, it is, of course, dynamic and changing" [35, p. 14].

Despite the great difficulties with text processing, CMD continues to expand. When the texts in the CMD appear, change, fade away for various reasons – this is exclusively by human design. These changes are ceaseless.

> About one page out of five is younger than eleven days. The mean age is around 100 days, so about half of the web content is younger than three months. The older half has a very long tail: about one page out of four is older than one year and sometimes much older than that. [42, p. 260]

The oldest page in the text collection mentioned above was dated 1992. Under the "age" of the web page, the period of time meant between the loading of the website and the date of its last modification. These statistics were obtained as a result of the observations of more than two million web pages, created by more than 25 thousand users. The observations were conducted over approximately a hundred thousand pages for seven months [42, p. 258].

Such data, received during the retrieving and mining of information in one way or another, is usually used after the secondary processing of it by researchers. It is not possible to rationally analyze the CMD but just place the received data in a linguistic "system of coordinates". This is the stage when the human mind operates by using heuristic tools. In this period the real "success" of program acting can be estimated and verified, changing data into ***knowledge***. In this way computer programs can be programmed to achieve the result but they have no “understanding” of what was achieved. Naturally, computer programs are oriented towards the stereotypes of human behavior and linguistic rules are really important, but the wider the CMD, the deeper the linguistic modeling of the semantics, or *information operating*, should be. Indeed, the information is not produced by a computer and that is why it cannot be “retrieved” directly from the computer. Information can be created by a person while carrying out intellectual coding and decoding of speech practice, which can be mediated or not, and data generalizing.

In such circumstances ***knowledge-basis*** will sometimes substitute CMC-sphere *data-basis*, leaving just supportive attributing functions to data. Data are transparent but simple. Human communication could not consist of ciphers and other primitive signs only: its essence is information, where quantums are *informathemes*, which could in turn be generalized into ***concepts***. The role of information as prima materia in CMC organization is key. It is growing as communication becomes more complicated, and is destined to be a kind of knowledge equivalent.

> The instance of the electric light may prove illuminating in this connection. The electric light is pure information. It is a medium without a message, as it were, unless it is used to spell out some verbal ad or name. This fact, characteristic of all media, means that the "content" of any medium is always another medium. The content of writing is speech, just as the written word is the content of print, and print is the content of the telegraph. If it is asked, "What is the content of speech?" it is necessary to say, "It is an actual process of thought, which is in itself nonverbal." …When the light is being used for brain surgery or night baseball is a matter of indifference. It could be argued that these activities are in some way the "content" of the electric light, since they could not exist without the electric light. [145, p. 8–9]

Of course, since the middle of the 20th century – the time of McLuhan’s writing – technology has changed a lot. *Electricity* seems a simple thing compared to modern inventions though its epoch is just over a hundred years. Of course, electricity has *always* existed but only recently has it been examined in depth. Information has a very similar history: it is an everlasting hypostasis of communication but only now do we have to know how to operate it consciously and entirely. It became “visible” in the scale of CMC, especially network-driven CMC.

Networking-specificity is the bright feature of current informationalization. Such functionality does not limit communication but widens it without contradicting the informational nature of speech. This vector of communication development was predicted and understandable already in the middle of the 20th century. One of the founders of the famous Artificial Intelligence Laboratory, or AI Lab, at Massachusetts Institute of Technology (MIT), the author of the *Frame Theory* Marvin Minsky, foresaw the creation of such a network well ahead of the creation of the World Wide Web in 1974.

> The frame systems are linked, in turn, by an *information retrieval network*. When a proposed frame cannot be made to fit reality – when we cannot find terminal assignments that suitably match its terminal marker conditions – this network provides a replacement frame. These interframe structures make possible other ways to represent knowledge about facts, analogies, and other information useful in understanding. [149, p. 113]

Minsky has even introduced a new term for such networks – *information retrieval networks*. However, Marvin Minsky was not the first and only scientist who saw the benefits of the informational potential of communication space. In one way or another, experts in many fields of science spoke about the urgency to arrange the amount of knowledge that was ever expanding due to technogenic inventions. In the evolutionary process, the names of the inventors of *literacy*, or *linguistics,* were lost, but this does not diminish their merits. In the same way, it looks like nobody knows who invented *information*, the importance of which is compatible with that of literacy or linguistics. New communicational capabilities allow the achieving of significant progress even in already well-studied areas. In reality, artificial languages and linguistic procedures inspired significant progress in all spheres of human life, including medicine, transport, economics, etc. No doubt that, step by step, the current informational "chaos" will be structured, described and modeled.

With that, informationalization has a wide range of limitations. The following *limitations of CMC informational continuum* can be considered linguistic or ***intralinguistic***: 1) formal character of mediating apparatus; 2) narrow range of paralanguage communicational means; 3) fragmentary linguistic instrumentality oriented towards the lexical level; 4) lack of contextual implementation of content that requires additional nature-language support and grounds in the traditional communicational environment (Fig. 1-1).

Thus, the semiotic factor is working here with *artificial languages* or formalized natural languages, proving "the formal character of the mediating apparatus". While solving some superficial problems of communication, such secondary semiotic systems create an additional linguistic dimension for metalinguistic interpretation and representation of communication. Improving computers creates an additional mediating "wall" between people and the computer. This mediation has a very clear task: achieve a heuristic-algorithmic consensus – though it does not work properly for now.

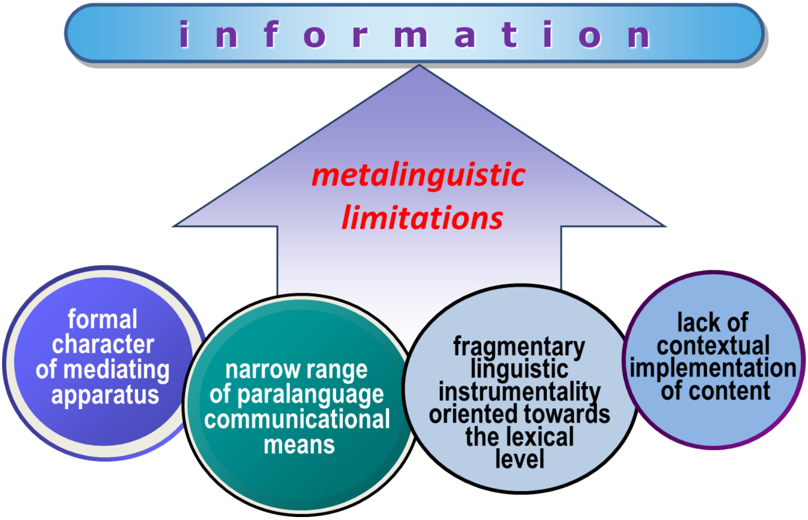
information
metalinguistic limitations
formal character of mediating apparatus
narrow range of paralanguage communicational means
fragmentary linguistic instrumentality oriented towards the lexical level
lack of contextual implementation of content

Fig. 1-1 – Limitations of CMC informational continuum

In addition, there are some ***extralinguistic*** limitations. One of the important *limitations* of computer-mediated activity is its narrow ***operationality***. Computer programs, operating with discourse, cannot take into account the entire discursive continuum. Theoretically this amount of knowledge could be formalized, too, but perhaps only after a hypothetical synchronization of a computer with the human brain. But the problem is the computer is not able to watch, analyze and, finally, generalize speech practice. In contrast, the human mind can and does – permanently and in parallel. The significance of such a feature is deciding – it is why waking up the next morning we can continue the affairs of yesterday, last year or last century. Similarly, a person can continue the work of other people or in another situation. Computer memory is not purposed for this. Moreover, it is too inefficient: computer capabilities are still limited *spatially* and *temporally*. The well-known example is of a *chess game*, which has been repeatedly cited by F. de Saussure, which is still controlled by computer programs to four or five following "steps". If a chess player knows what usually happens after, let's say, the fifth move, he wins easily. He cannot calculate it but he should know. No doubt, tomorrow's computer can be programmed to the “sixth” move; eventually the whole chess game will be totally described for the computer. But this could be easier to model than calculate, and it can hardly be suggested without effective modeling.

The partial ***incompatibility*** of speech sense with ambiguous “fuzzy” data of language is another vital limitation of informationalization. It can be overcome only in human “neural” mode. With that, *neural networks*, *deep learning* and other improvements are still just commercial metaphors based on the old statistical model means. Speech dynamics is a relevant limitation here as well: the ***actuality*** of sense differs for the different attempts to freeze it in meta-descriptions. Moreover, the speech practice of CMD is an example of not only a very extensive database but also a ***multilevel*** system: many sociocultural circumstances present and work inside any text. In CMC there is clearly a new degree of influence on the informationality of language practice – the influence of ***technogenic*** factors. In many ways such influence is mediated via the technological means of the new generation.

The coding format used today in CMC is characterized by the many *limitations* of informationalization properties. These restrictions could be “loosened” or diminished either by improving the statistical model or by substituting the model itself. A third way is the creation of a new kind of human, whose mind will be adjusted to be a computer – fortunately, that hasn’t happened yet. Of course, it is easy to teach people to think like a computer: every computer owner acts like this sometimes. But after that, they can switch from algorithmic to heuristic thinking and solve many other problems with their human brain.

1.4 Linguistic reincarnation of information: Linguoinformationality

CMC-practice demonstrates the increased variability of its textual content. Of course, some part of CMD is represented by copies of texts already published in traditional ways: copies of copies, variants, cover versions, etc. Undoubtedly, due to the possibilities of speech, material comprehension and processing speed, the computer support of electronic texts is much more effective than traditional handwritten and analogue meta-language means. With all the superficiality of the algorithms of the CMC-sphere that are used for text *creation*, *modification* and *usage*, an important advantage is their high technical precision. Such an advantage stimulates the total translation of printed and multimedia texts into a computer-mediated format. Artificial intelligence programs follow an