Journal of Siberian Federal University. Humanities & Social Sciences 1 (2013 6) 50-61
~ ~ ~

УДК 81'322; 004.934; 004.912

Philosophical and Theoretical Background on Development of Text Corpora

Alekseiy Yu. Mordovin*
Irkutsk State Linguistic University
8 Lenina St., Irkutsk, 664025 Russia [1]

Received 09.01.2013, received in revised form 16.01.2013, accepted 23.01.2013

The publication is devoted to the philosophical and theoretical background on development of the first computer-based text corpora in 1960s. The key methodological features of the corpus-based approach to language analysis and the key corpus functions are researched, including early approaches to corpus method, the young grammarians' platform, the role of the corpus approach in structuralism, contribution of descriptive linguistics into the development of corpus method. This analysis tracks the development of prerequisites for the modern concept of the role of corpora in linguistic analysis throughout main text-centered linguistic theories encompassing the period from mid-XVII to mid-XX centuries. The article infers that apart from the apparent objective factors determining the development of the corpus method; all necessary theoretical preconditions for implementation of pro-corpus approach in the form of machine-based corpora were developing gradually.

Keywords: text corpus, corpus linguistics, history of linguistics.

Introduction

The rationale for analysis of the philosophical and historical background on development of text corpora and corpus linguistics is driven by the increasingly recognized awareness of computational linguists that once taken departure from "formal linguistics" needs to be revisited. At a particular moment in the past, computational linguists "frustrated with the dominating theories in formal linguistics, looked instead to the corpora that reflect language use as our sources of (implicit) knowledge" (Wintner, 2009).

For the past twenty years computational linguistics has evolved into "natural language engineering" due to being able to discover "the right mathematics". And now it seems that "there's much new in the world of linguistics, much that should interest us computational linguists" (Wintner, 2009).

With respect to this an expected effort would be to identify the new knowledge that formal linguistics can extend to assist computational linguistics. This paper, however, aims in the opposite direction. It undertakes to reconstruct the evolution of formal linguistics to a state, where it has given birth to corpus linguistics. In the later course of research, this knowledge could be useful in giving a more profound methodological insight into what caused the "frustration" about 25 years later.

The key factor, which triggered modern development of corpora and corpus linguistics,

* Corresponding author E-mail address: alexmordovin@mail.ru

was IT revolution. This is based on the assumption that without an efficient computerized search engine for the corpus, its maximum size, usability and, most importantly, its openness to infinite development of new utilization methods would be seriously undermined. Nevertheless, the IT factor being of technical nature is external to language itself.

Leaving IT matters aside, the article sets out to consider philosophical and historical pre-conditions for development of the main functions of text corpus, as well as to provide their first rough description.

This way the study could be described as overall methodological. Specific linguistic theories were selected for analysis in the paper based on the comprehension that the doctrine of text priority had been governing the development of western linguistics since the beginning of Christianity. The prerequisites against which the first computer-based text corpora appeared will be traced within the framework of theories commonly defined as: universal grammar, classical comparative-historical (comparative) linguistics, neogrammarians, structuralism (including its branches – glossematics and descriptive linguistics). One must admit that due to the constraints of the paper size, the range of scholars being quoted may appear somewhat arbitrary, and despite all efforts the author may undertake, the manner of presentation may remain fragmentary. Therefore, specific names mentioned in the present article do not necessarily appear as pivotal points, but rather than as an illustration of the respective epoch's common opinion.

In parallel with the text-centered block, pro-corpus approach was also evolving within the anthropological part. However, due to extensive differences in underlying rationales of the two parts, analysis of the anthropological part may not be included into the present paper, and will become the topic of a separate article.

The temporal framework of the period in question lies between the two events: completion of Reformation in Europe (Peace of Westphalia made in 1658) – as the starting date, and appearance of the first computer computer-based text corpus (The Brown University Standard Corpus of Present-Day American English – 1963) – as the ending date.

Early approaches to pro-corpus method

The starting point of this analysis will be the Port-Royal Grammar (1660). This early theory demonstrates both text-centered and human-centered attributes. However, it is not of interest per se, but as a vivid example of the rationalistic approach to language, which was dominating in European philosophy of XVII century.

The essence of the Grammar (Arno, 1998) is as follows: all people think in similar ways, therefore, building a grammar of any particular language is a useless effort. Language reflects thinking of a person, while logical structure of thought is common for all people/languages, and it does not equal to structure of any particular language. It should be noted that the way a scholar answers the question whether or not language is independent from thinking is an important methodological prerequisite, which entails yet another prerequisite – whether language is (in) finite. In many respects, the combination of answers to these two questions determines the status of corpora and pro-corpus approach in general within any specific linguistic theory.

The attempt to provide ourselves with an answer whether or not a text corpus would be appropriate within the theory of Universal Grammar brings us to two conclusions.

Firstly, it is known that this grammar should be considered a logical, or a priori grammar. In this respect, it is self-sufficient, and may seek a substantiation of its individual statements

among facts of any language. Therefore, a text corpus may be used within this grammar for the purpose of illustration. This function of text corpus remains one of the most demanded up to the present day. A text corpus may be loyal to most of a priori formulated doctrines, and if the search parameters are selected correctly, it will generate a virtually unlimited number of the necessary examples from speech. In other words, in return to a particular entry of search categories, a corpus will not normally produce what could be perceived as a negative answer, but will only reflect the quantitative expression of the probability (frequency), at which the language phenomenon in question occurs. This statement, in its turn, can be interpreted rather liberally.

On the other side, the idea that thinking is not guided or determined by language, which was proclaimed in the Grammar of Port-Royal (and much later – in generative linguistics of N.Chomsky), discredits the evidential potential of the corpus in the generation of new knowledge. Let us call the function of allowing generating new knowledge on the basis of analyzing the material systematized in a corpus the heuristic function of the corpus.

Thus, it appears that an empirical approach underlying all of corpus linguistics may not function as a reliable method of generating new knowledge within the framework of rationalistic doctrine. Nevertheless, deductive structures of universal logical categories driven by the need to have their own inference about language facts may resort to empirical experience of the corpus, while the latter, being a non-categorical object, readily allows for such application. There are sufficient grounds to state that no interest in pro-corpus approach as a tool of perception existed at the stage of linguistics development being considered.

The range of problems studied by comparative linguistics at a later time included matters of language genealogy, typology, stages and laws of language development, existence and extinction. Ancient written texts were the predominant material of the research. Greatest emphasis was laid on in research of the sounds, while it is exactly the level of language, where pro-corpus approach is applicable the least. Word morphology was more frequently studied in detachment from the meaning, and a comparison of text corpora in different languages could have been useful here, but the nature of research itself would have raised impossible demands to such corpus.

As a matter of fact, one must possess a specific set of grammatical categories and their markers, as well as a vocabulary, to units of which these categories will apply, in order to provide morphological mark-up of the corpus, which may later be used in an algorithm-based analysis and may produce interesting results leading up to new knowledge (heuristic function). Once the above tool is available, corpus contents must be marked-up with a purpose-designed software – fast and automatically, or manually – slowly and labor-intensively.

Meanwhile, in the course of a comparative study of language, it was precisely the analysis of morphological parameters which comprised the scope of study, while the result was often seen as a compendium of the discovered language facts, and less often – inference made on its basis related to typology, genealogy or stages of language development, and less seldom yet – objective laws of language development. It is hardly worth mentioning that any language laws obtained by means of this much labor were suffering from numerous exceptions.

Philosophy of the young grammarians as theoretical recognition of pro-corpus method

Positivism as the new philosophy sprouted the new ideology of the young grammarians. They

set aside the attempts to resolve any of "eternal" matters, which would not be supported by facts. Quite similar to the pro-corpus approach, the mission of the research was seen in observation, recording and consolidation of facts. Consolidation was still understood as restoration of particular proto-forms; however, the range of methods used was substantially expanded in a corpus-favorable way: historical and psychological data were included, and some experimental work had begun. Modern stage of language existence was finally relieved from perceived decay and aging, but most importantly – the priority of live and accessible language material as a departing point for all research were proclaimed.

At the stage of young grammarians' theories, the future pro-corpus approach was first admitted to be capable of generating new knowledge – the second most important function of corpus named in this paper the heuristic function. This function is exercised whenever language material organized in a corpus-type body is not only used to illustrate a particular a priori theory, but it also retains the right to be a source of new, undiscovered knowledge, subject to consistent application of research methodology.

Meanwhile, as consistency and reliability of implementation of language facts in the research material was growing, the status of generalizations obtained as the result of their research was growing on a pro rata basis. The notion of sound law in language development appeared as a well-expectable outcome of this growth. It can be said that linguistics was gradually preparing for future convergence with scientific and mathematic methods of research.

In addition, after the young grammarians rejected the metaphysical "national genius" of W.von Humboldt, and assumed psychologism as a principle, which explains the collective nature of language, they strengthened the future position of text corpora even more. H.Paul stated that we must recognize, properly speaking, that there are as many separate languages in the world, as there are individuals (Paul, 1960). From then on, when striving for a truly complete recording of language facts for the purpose of their future consolidation and generation of language laws, the only finite limit should have been description of the language spoken by any individual.

Nevertheless, from the point of view of psychologism any individual's language does not boil down to integrity of all texts generated by this individual, but rather to "an integrity of representations confined within his soul, which belong to speech activity". In order to describe truly collective aspects of language, H. Paul introduces the notion of usage. Therefore, in strict sense, today's corpora are filled with precise usage – which is a secondary collective product of the truly real psychological constructs of the individual's mind.

H. Paul warned linguists to avoid indulging into abstractions and to take their time with any conclusions. Instead, practical collecting of material was safer: "When we unite the languages of numerous individuals into a single group, and set it against languages of other individuals... then we always sidetrack ourselves from one kind of differences, but consider others. There are a lot of opportunities here for abuse of discretion".

In strict methodological sense, all resort of H. Paul and the young grammarians in general to psychologism as a final reality is not a systemic phenomenon for the young grammarians. More likely, it is an attempt to conceptualize the problem of structuring meaning in general, which could be easily avoided for the time being, while phonetic or morphological objects were placed under analysis, but kept recurring in the word analysis. Later, this problem resulted in a number of theories describing various "layers" in the meaning of a word, and then the issue

of meaning became ousted into the domain of speech linguistics.

Still, the young grammarians' approach cultivates the degree of research rigidity, which was unmatched before. It asserts consistency and verifiability of the results – the qualities so clearly manifested in the pro-corpus approach, which was yet to come.

The need to analyze live language to generate inductively new knowledge continues to develop among linguists of transient period between the young grammarians and F.de Saussure. For instance, I.A. Baudouin de Courtenay (Sharadzenidze, 1980) stated the need for a "comprehensive analysis of positive data in already existing languages", i.e. "live languages in all their diversity".

Overall, it should be understood that the concept of abstraction often functioned as the prototype of corpus linguistics, even though the heuristic value of abstraction was receiving very different appraisals. The so-called statistic text corpus and any results, which may be obtained on its basis, are nothing but an abstraction. An individual's language and abstraction contained in corpus are controversial, but this controversy is productive – that is the point stated by linguistics adhering to psychologism. Following ideas of the young grammarians, I.A. Baudouin de Courtenay agrees that the only true reality of the language of an individual. It is the only entity that does not deserve to be called a mere abstraction, since the processes occurring in human brain are real. As for the rest of "languages" – whether it be Polish or Russian – they are already a particular degree of abstraction.

Eventually, a respectable text corpus is exactly this kind of abstraction, which has been evaluated positively. This is why the idea of the language as an abstraction was not only preparing the ground for sprouting the ideas of structuralism, but also was creating preparedness for abstraction from the undoubtedly real language of an individual toward a less-than-useless abstraction of a finite size.

A genius linguist, I.A.Baudouin de Courtenay was already able to foresee at the junction of XIX and XX centuries that linguistics will gradually "mathematize" by developing methods of quantitative (including statistical) and qualitative analysis. That is to say, he believed in conclusiveness of abstractions provided that the young grammarians-style research stringency is ensured, while the research is based on the real "bricks" of individual idiolects.

Instrumental role of pro-corpus method in structuralism

The main stage of development of text corpora to the present state and their functioning as both research material and a tool for linguistic research occurred during the epoch of structuralism. In the theory of F.de Saussure itself (Saussure, 2004), language occupies the central place as a systemic semiotic object. The contents and affiliation of text corpora, according to Saussure's theory, should undoubtedly be referred to speech. However, it is well known that as opposed to linguistic of language, Saussure's theory lacked a clear-cut definition of speech linguistics.

But his nearest successor (and co-author) Antoine Meillet emphasizes social nature of language rather than psychological nature stated by the young grammarians: "Language, being, on the one hand, a possession of individuals, is, on the other hand, imposed onto them; thanks to that it is a reality of not only physiological or psychic type, but most importantly – of social type".

In the above quotation it is fairly easy to observe a number of vital paradigm shifts, which are extremely favorable for the text corpora: from reluctant recognition of language as a useful abstraction from a number of real phenomena of

psychological genesis, the concept of language grows to a "social reality", which is of relatively external nature to an individual. This reality, as opposed to abstraction, may be studied by means of pro-corpus methods with greater assurance. The key difference of understanding text corpus in structuralism versus the young grammarians is the detachability of language from the individual.

When one sets out to study language in its structural appreciation, it is no longer hazardous that the researcher may admit a too high level of abstraction from the individual language reality, nor it is advisable to carefully accumulate language material so as to ensure maximum accuracy of language law. On the contrary, the time has come to liberate language from individual idiosyncrasies as much as possible, to strip its structural socially-determined essence. It is here that depersonalized (i.e. systemic) and statistic (and therefore – truly social) pro-corpus approach becomes appropriate as never before.

For the first time, starting with structuralism, the pro-corpus approach is assigned to the role of primary collecting language facts, which are then used in order to build a synchronous slice of the language system. Much later, a version of this text corpus function was developed independently in the form of preserving the maximum possible number of texts in disappearing languages as text corpora. This function of corpus is called the preserving function.

Despite absence of complete agreement among structuralists concerning distribution of synchrony and diachrony status in language, a corpus is always a more or less synchronous slice, while any pair of corpora with all remaining parameters equal save time is a material for diachronic analysis. Thus, a new metaphor evolved in structuralism to denote text corpora – it can be viewed as a virtually perfect "slice" of language synchrony, and, accordingly, becomes a very valuable object of research as opposed to ever-changing real language.

This gives rise to the new question: while some corpora are designed with fixed time frame for texts to be included. Meanwhile, some of the largest national text corpora are designed without an explicit tie-in to a time interval, or this interval is of considerable duration. In this respect, the issue of status of any given national corpus in general and of a particular corpus in question must be decided. Depending on the decision made, several action plans could be proposed:

1) identifying text corpus as a contemporary corpus by providing a timespan. Later, "outdated" texts must be removed from the corpus promptly;

2) identifying text corpus as "supertemporal", or establish a very broad time frame (which is, basically, the same);

3) exclude the issue of synchrony/diachrony for the given corpus, by delimiting the corpus from the rest of the language with non-temporal (perhaps non-linguistic) framework.

In the latter case, when corpus integrity is not justified either by its belonging to synchrony, or to diachrony of a particular time span, it is assumed that integrity and consistency of corpus are being derived from the primary object, which the corpus is designed to reflect – from language itself. In general, language is an out-of-time entity, which retains its integrity regardless of any occurring change. It appears, that this representation of language, as manifested in text corpus, is supported by the external, transcendental nature of the "language cause", such as "genius" for W.von Humboldt (Humboldt, 2000) etc.

Something of similar type related to structure of language as one of the aspects of its integrity and consistency (for a structuralist researcher – only at a given moment in time), can also be observed in structural linguistics.

In particular, in Viggo Brøndal's opinion (Zvegintsev 1964), the value of structural linguistics, among others, lies in the fact that in structural linguistics the overly positivist approach of XIX century linguistics becomes more realistic thanks to "legalization" of deduction and emphasis on structural nature of language phenomena. Pure induction does not exist; it is simply a necessary instrumental stage, which precedes the deductive system-building. Experience rests on hypotheses, on fundamentals of analysis, abstraction and generalization; therefore, induction is nothing but a masked deduction. In Brøndal's opinion, the structure is typical of all real phenomena, not only language, and it is this particular quality, which ensures integrity of all objects.

In this case, corpus is an object perceived as a sufficiently accurate model of integral structured object – language, and it "inherits" its integrity and structure, being entitled to function as the object of language research.

Albert Sechehaye's linguistic theory of organized speech (Sechehaye, 2003) points directly at the intermediary position of real speech (i.e. contents of a corpus) between synchrony and diachrony, which is of vital importance for linguistics: speech is based on a particular state of language, therefore it is an element of synchrony, however, it already contains "seeds" of all possible changes even before they occur, and therefore belongs to diachrony.

This approach makes generation of new knowledge by means of systematized analysis of large volume of speech material especially justifiable and feasible, since it allows captivating qualitatively and quantitatively the moment when changes appear in synchrony, which is bound to constitute a recording of a diachronic fact. In real world, a linguist will be faced with the task of analyzing specific occasional phenomena: selection of language units, matters of style etc. a large statistic corpus will be required to this end. A linguist will be enabled to research facts from the point of view of instrumental (functional) positivism, in a fast and efficient manner, without stalling in analysis itself due to the available technical capacities. Then, he will be able to shift to inductive-deductive conceptualization of integral systemic aspects underlying the detected change.

From the methodological point of view, such research is a combination of instrumental positivism followed by deductive idealism, which was insisted upon, among others, by Karl Vossler (Vossler, 2007). Moreover, automation and acceleration experienced the greatest demand ever particularly in the linguistics of organized speech. By definition, this linguistics' object is a large volume of language facts, only part of which will be accepted by the language community and will migrate into the domain of diachronic linguistic facts, while the remainder will be rejected and will not become developed. Corpus linguistics, in this case, can only be a perfect tool for linguistics of organized speech: it does not renounce the diachronical studies, but only prepares, or "concentrates" language facts for the latter, while the comparative scholar is left to decide whether or not they are systemic.

As we see, in Albert Sechehaye's theory corpus linguistics already acquires an explicit methodological purpose and its status is being defined.

Glossematics of L. Elmslev (Elmslev, 2005) from the modern standpoint is often perceived as the first and quite successful attempt to introduce mathematical stringency into the language science. This was accomplished at the expense of very significant reduction and depletion of its object. Despite the glossematics' pro-Humboltdian theoretic preface, no other linguistic theory has ever departed from the

speaking human as consistently as glossematics did. For purposes of the present paper, it would be reasonable to consider the causes, which drove L.Elmslev to build a general language theory by consistently applying the deductive method, on the most common principles borrowed from mathematical logic.

The main cause is apparent – to relieve description of language from a shade of contradiction, to make it exhaustive and ultimately plain. This theory then has to be independent from experience. It determines its object independently by means of an arbitrary and suitable selection of prerequisites. Maximum the practical material can count on is that the theory will be constructed arbitrarily in such manner, which would be applicable to describe objects of particular nature in a non-contradictory and exhaustive way. Incidentally, experimental data will never be able to strengthen or attenuate the theory; they may only strengthen or attenuate its applicability.

This position, at first sight, appears to be an unconditional refusal from any empirical study undertaken to collect new data on the basis of analyzing a large bulk of material.

However, does the author indeed object to text processing? On the contrary, the author welcomes it as a primary stage of accumulating knowledge collected inductively. The scholar agrees that using a tool set of the linguistic theory, a stock of knowledge can be extracted from a selection of texts (in modern terms – from corpora), which can then be utilized in other texts. Abiding by the principle of text priority, the author does not refuse the right to exist for the instrumental deduction comprised of primary material accumulation. Despite the fact that the theory will then be built not on the basis of facts collected, but quite the opposite, facts will only comprise a complementary approximate guideline for applicability of the theory, Elmslev believes that operational collection of data on the basis of a "selection of texts" is of ultimate importance.

In other words, Elmslev believes it is necessary to engage the heuristic function of pro-corpus approach at the initial stage to select the design target of the theory, while later, he may be prepared to utilize the illustrative function in order to resolve the practical mission of describing the language by applying the theory to a particular material. The specific feature of his approach lies in the fact that he refuses to make use of yet another function of text corpora. Let us call this function the verifying function. This function explains the ability of text corpus to function as a method of semi-quantitative assessment of theory accuracy.

The above considerations allow for a number of interim conclusions:

1) F.de Saussure's theory lacks a clear definition of status and role of speech linguistics. Nevertheless, this should not be mistakingly taken for a refusal from pro-corpus method of linguistic research as opposed to the young grammarians' theories. Also, this is the beginning of rethinking the status of pro-corpus approach (from basic to auxiliary).

2) In subsequent more detailed theories of structural linguistics pro-corpus approach was consistently being given the role of an obvious primary stage of gathering a particular knowledge, which could be used for systemic conceptualization in the future course of research. The method of utilization of the obtained knowledge may differ, which does not affect the demand for the method.

3) The notion of the role assigned to pro-corpus approach within the framework of structural linguistics is in compliant with the ideology of acceptable instrumental positivist-like collection of language facts in the Humboldtian ideology. In both cases, collection of facts is not

the final goal of the research, but the nature of their application is expectedly different.

No fundamental distinctions may be observed in underlying principles for development of pro-corpus approach among other scholars, who wrote within the framework of the structural approach, including in line with I.A. Baudouin de Courtenay's tradition.

L.V.Scherba (Scherba, 2004) specifically states that language material is the integrity of all spoken or comprehended in a particular setting during a particular epoch in life of a given social group. Speaking in linguists' language, we are talking about texts. Language material is the result of language activity. Texts include language units, and all language units, which we operate within a dictionary or grammar, are concepts, and may not be observed in any immediate experience (either psychological, or physiological), they can only be extracted from language material.

In accordance with this, language system is a derivative from language material, it is "objectively embedded in the given language material and manifested in "individual speech systems", which may arise under the effect of this language material". Meanwhile, the warrant of the integrity of language system is the integrity of language material within the framework of a particular social group (i.e. in the so-called "organic" text corpus). "Linguists are entirely correct, when they derive language system from respective texts, i.e. from the respective language material. Language system is not stated arbitrarily, but in accordance with the information extracted from language material".

In addition, L.V.Scherba recognizes the need for linguistic experiment with the purpose of verifying data collected as the result of texts analysis against *more diverse material*, thus attracting new contexts into the analysis. We may see that the verifying function of the text corpus is recognized as quite acceptable.

In certain cases if a theory belonged to structuralism, it did not preclude the opportunity to apply pro-corpus approach at a higher than usual level, namely, not only to prepare language facts as a material for deduction, but also to justify the selected option for the deductive structure of language system.

This was possible whenever language was not diluted to a system of pure relations, but would include the social factor, for example, as in the theory of Jerzy Kuryłowicz (Zvegintsev, 1960), where the domain of sign application within the system corresponds to the domain of its application in a language community. In other words, the more general is the contents of the sign, the wider is the area of its application by speakers; the more specific is the contents of the sign, the narrower is the area of its application, not only internally, but also externally. As we see here the quantitative results of applying pro-corpus approach are leading to the immediate statements concerning balance of system.

Descriptive linguistics as theoretical prototype of pro-corpus method

The author of this paper believes that the first computer-based text corpora, which appeared in 1960s, can best be interpreted according to the ideology of descriptive linguistics. While preserving fundamental theoretical views of structuralism, representatives of descriptive linguistics also introduced a number of important methodological aspects, which have completed the general theoretical background for the beginning of the pro-corpus method.

The descriptive linguistics arose out of anthropological studies of the aboriginal peoples of the USA and Canada. As structural theories of language had reached a particular level of maturity, linguists were forced to digress from all elements of personal introspection, which

virtually excluded the effect of the European languages onto the research. With regard to the aboriginal languages they were studying, the system of written language was yet to be developed, while major difficulties delimitating the words, isolation of meaning etc. as compared to European languages. Due to conceptualization of the theoretical implications of using bilingual informants, descriptive linguists were able to model the future scheme of interaction with a text corpus.

Whenever the informant was asked questions whether a particular language distinction is or is not meaningful, his/her responses were naturally based on his/her own language intuition. It may seem as if the function of introspection has simply been passed on from the researcher to the informant. However, the fact of greater methodological value is that when the researcher was looking for answers to questions about the structure of language, descriptive linguistics build on the principle of "black box". That is to say that the nature of interaction between the researcher and the informant was quite similar to "human-machine" interface. Special emphasis was laid on how to formulate questions correctly, and, generally speaking, most of field linguistics methods were developed at the time.

This experiment became the prototype of interacting with the informants' own language (a European language) as with a "black box". So the formal, mechanized, statistic approach to researching of the language was gradually coming into practice. Descriptive linguistics took every effort to drive any subjective elements out of linguistic analysis using a number of methods, which were bringing it closer to the pro-corpus approach.

The procedures of segmentation and distribution developed by them became the prototype of modern corpus mark-up. Also, it is of value that as methods of distribution analysis were developing (which is a purely mechanistic procedure, not based on meaning), descriptive linguistics was anticipating of avoiding the need to use informants altogether. Any kind of intuition had to be replaced by quite "machine-based" principles relying on mathematical methods: procedural approach to language, criteria of text segmentation, distribution analysis, authentication of units based on distribution.

In fact, after the first machine-based corpora of the English language appeared, the situation of linguistic research has changed the performers, but retained their roles. Now the search interface of the corpus was functioning as the informant. "Communication" with this new informant took place in accordance with a set of strictly formal criteria, while the object of research was not a totally unknown language of an aboriginal tribe, but the researcher's own language. The need for intuition and introspection was phased out by representativeness of corpus and statistic support of the results, while the internal structure of this "black box" could at last be studied using methods of mathematical nature. Undoubtedly, the overriding role in the process was played by the forced interest of descriptive linguistics to syntagmatics of language versus its paradigm, since it is particularly the research of syntagmatic language parameters, which is the most feasible using text corpora.

Finally, after theoretical and methodological prerequisites for language research with pro-corpus, methods were developed to their logical limit within descriptive linguistics, R. Jacobson (Jacobson, 1985) was able to point at yet one more unexplored area, where text corpora could be applied. He called for the re-discovery of the problems of language universalia by stating that "only now does linguistic have at its disposal all methodological prerequisites required to design an adequate model". What were these prerequisites? He meant that alongside with the development

of differential attributes theory in phonology, it also became a versatile system for differentiation of symbols at other levels of language. However, when a developed common system of describing them was lacking, universalia used to be isolated completely and inductively, following the results of empirical analysis of materials over numerous languages. Using a "fast" pro-corpus methodology over a large number of languages could produce most unexpected and hasty results in search of language universalia.

Moreover, the time was coming to conceptualize the relations of a human and a computer in the domain of natural language, which drove R. Jacobson to search for parallels between linguistics and theory of communication. The idea of language communication as coding and decoding of information itself (on the side of the speaker and listener, respectively) became extremely useful and appropriate to justify the pro-corpus approach. As such, the research of language facts based on a statistic corpus is a transformed act of decoding, when instead of the listener, whose mechanisms of perception are encumbered with hard-to-replicate cultural, individual, subconscious parameters, otherwise "inconvenient" for externalization, the impartial machine takes its place to "listen" to speech with a set of pre-developed special algorithms. Already then popular attempts to combine linguistics with mathematics if not by means of specific methods of research, then at least by means of a common methodological perspective, made this idea so favorable that even the problem of meaning stopped being considered as critical as before.

Conclusion

Thus, the performed analysis of the historical and philosophical prerequisites for the appearance of text corpora allows drawing the following conclusions.

When researching factors, which determined appearance of the first machine-based corpora, it is possible to build on the development of IT being a technical cause for the beginning of corpora. Analyzing the key text-centered trends of linguistics over the period of about three centuries up to the beginning of the first corpora allows inferring that all necessary theoretical preconditions for implementation of the pro-corpus approach in the form of machine-based corpora were developing gradually. As linguistic ideas were evolving, key methodological and structural elements of corpus linguistics were anticipated on a number of occasions. Moreover, by the actual moment when text corpora began, one could assert possibility of a developed concept of their role in the structure of linguistic research.

References

1. Arno A. *Grammatika obshchaia i ratsional'naia Por-Roaialia* [General and Rational Grammar of Port-Royale]. Moscow, Progress, 1998. 272 p.
2. Elmslev L. *Prolegomeny k teorii iazyka* [Prolegomens to Theory of Language]. Moscow, URSS, 2005. 248 p.
3. Humboldt W. von *Izbrannye trudy po iazykoznaniiu* [Selected Writings in Philology]. Moscow, Progress, 2000. 400 p.
4. Iakobson R. *Izbrannye raboty* [Selected Writings]. Moscow, Progress, 1985. 460 p.
5. Meillet A. *Sravnitel'nyi metod v istoricheskom iazykoznanii* [Comparative Method in Historical Philology]. Moscow, URSS, 2004. 108 p.
6. Paul H. *Printsipy istorii iazyka* [Principles of the History of Language], Moscow, Foreign Literature Press, 1960. 501 p.

7. Saussure F. de *Kurs obshchei lingvistiki* [Course of General Linguistics], Moscow, URSS, 2004. 278 p.
8. Scherba L.V. *Iazykovaia sistema i rechevaia deiatel'nost'* [Language and Speech Activity]. Moscow. URSS. 2004. 432 p.
9. Sechehaye A. *Programma i metody teoreticheskoi lingvistiki* [Program and Methods of Theoretical Linguistics]. Moscow, URSS, 2003. 264 p.
10. Sharadzenidze T. *Lingvisticheskaia teoriia I.A. Boduena de Kurtene i ee mesto v iazykoznanii XIX-XX vekov* [Linguistic Theory of I.A. Baudouin de Courtenay and its Place in Philology of XIX-XX centuries], Moscow, Nauka, 1980. 136 p.
11. Vossler K. *Esteticheskii idealizm. Izbrannye raboty po iazykoznaniiu* [Esthetic Idealism: Selected Writings in Philology]. Moscow, LKI, 2007. 144 p.
12. Wintner S. What Science Underlies Natural Language Engineering? *Computational Linguistics December 2009*, Vol. 35, No. 4: 641–644.
13. Zvegintsev V.A. *Istoriia iazykoznaniia XIX-XX vekov v ocherkakh i izvlecheniiakh* [History of Philology of XIX-XX Centuries in Essays and Extracts]. Moscow, Prosvescheniye, 1964. 466 p.

О философско-исторических предпосылках к появлению корпусов текстов

А.Ю. Мордовин
Иркутский государственный лингвистический университет
Россия 664025, Иркутск, ул. Ленина, 8

В статье рассматриваются философские и исторические предпосылки к возникновению первых компьютерных корпусов в 1960-х годах. Рассматриваются основные методологические характеристики корпусного подхода к исследованию языка и функции корпусов текстов. Развитие предпосылок к современному пониманию роли корпусов текстов в лингвистическом анализе прослеживается в основных текстоцентрических направлениях языкознания за период с середины XVII до середины XX века.

Ключевые слова: корпус текстов, корпусная лингвистика, история языкознания.