# CORPUS LINGUISTICS APPROACH TO NATURAL LANGUAGE INVESTIGATION

Rubashko Natallia

Faculty of Applied Mathematics and Informatics, Belarussian State University, Skorina av., 4,
Minsk, 220050, Republic of Belarus, e-mail: *Roubashko@bsu.by*

**Abstract.** The article analyzes the revival of corpus linguistics approach to natural language (NL) investigation concerning the solution to the following problem: how to provide a computer with the linguistic sophistication necessary to perform robust, large-scale natural language processing (NLP).

## 1. Introduction

The development of a NLP system must handle the following problems [15]:

- *knowledge representation*: how to organize and describe linguistic knowledge in a linguistically meaningful and computationally feasible manner;
- *knowledge control*: how to apply linguistic knowledge for effective processing;
- *knowledge integration*: how to use the various knowledge sources effectively;
- *knowledge acquisition*: how systematically and cost-effectively to acquire the required knowledge and maintain consistency of the knowledge base.

One of the biggest challenges in NLP is overcoming the bottleneck of linguistic knowledge acquisition (due to its important role): solving large-scale, real-world language processing problems in a robust and predictable way. These problems include information extraction, text summarization, document generation, machine translation, second language learning, amongst others [2, 3, 4, 8]. In many cases, the technologies being developed are assisting, rather than fully automatic, aiming to enhance or supplement a human's expertise rather than attempting to replace it. Two features of the current situation are of prime importance. First, there is no theory of language which is universally accepted, and no computational model of even a part of the process of language understanding which stands uncontested. Second, building intelligent application systems, systems which model or reproduce enough human language processing capability to be useful, is a large-scale engineering effort which, given political and economic realities, must rely on the efforts of many small groups of researchers, spatially and temporally distributed, with no collaborative master plan [8].

There exist many NL investigation strategies suggesting the solution of problems mentioned above. All strategies can be divided into groups of approaches closely connected with two scientific directions: artificial intelligence (AI) and computational linguistics. The primary goals of AI are the development of computational methods for NL understanding. The goals of computational linguistics are the processing of NL close to human performance [13].

The first approach is based on the assumption that a computing system for successful NLP should be able to use vast knowledge resources and make logical deduction on the basis of this knowledge. This approach called *traditional, rationalist* [2], *knowledge-based, rule-based* [15], *AI approach* [13] had been the dominant paradigm in the natural language processing community. This approach usually has the following characteristics [15]:

- the linguistic knowledge is usually expressed in terms of high level constructions such as parts of speech (POS), phrases, syntactic trees and feature structures described in most traditional linguistics textbooks, and the knowledge is expressed in the form of syntactic or semantic constraints over such constructions;
- most rule-based systems have a strict sense of well-formedness; therefore, the rules are applied deterministically to reject ill-formed constructions;
- most rules are based on existing linguistic theories that are linguistically interesting. When the required knowledge does not appear in the literature, ad hoc heuristic rules may be used;
- rules are normally induced by linguists based on their expertise.

Such approach, however, suffers from serious difficulties in knowledge acquisition in terms of cost and consistency.

The second approach, with the capability of automatically acquiring knowledge from real text data, is becoming more and more popular, in part, to amend the shortcomings of AI approach. It is called *empirical, corpus-based* [2], *statistical* [15], *linguistics engineering approach* [12] later developed in methods of reproducing simulation of linguistics objects and processes [13]. It has the following characteristics [15]:

- the knowledge can be described using a set of features and the associated probabilities in statistical approaches;
- there is not a strict sense of well-formedness. Therefore, a statistics-oriented system may adopt a statistical language model by using likelihood measures for choosing the most likely analysis among all;
- statistical approach would automatically acquire the knowledge, which is the probability values in this case, via estimation processes, from a corpus.

The empirical approach is much more data driven and is at least partially automated by using statistical or machine-learning methods to train systems on large amounts of real language data: a big collection of naturally occurring sentences which is called a corpus. In computational linguistics it was formed specific scientific direction – *corpus linguistics (CL)* [10].

The increasing availability of machine-readable corpora has suggested new methods of NL investigation. Though common in the speech community, the use of statistical and probabilistic methods to discover and organize data is relatively new to the field at large. The various initiatives currently under way to locate and collect machine-readable corpora have recognized the potential of using this data and are working toward making these materials available to the research community. Given the growing interest in corpus studies, it seems timely to devote an issue of corpus linguistics to this topic.

## 2. From the history of corpus linguistics

The development of methods of NL investigation was influenced by the research in AI almost to the end of the 1980s. Since the automatic NL processing and its theoretical basis – structural-mathematical linguistics – were born and formed during the period when leading linguists took a great interest in the idea of discrete-atomic structure of language [11] it was resulted in distributing and strengthening the opinion that only AI-methodology was eligible for NLP.

The 1990s have witnessed a resurgence of interest in 1950s-style empirical and statistical methods of language analysis. Empiricism was at its peak in the 1950s, dominating a broad set of fields ranging from psychology (behaviorism) to electrical engineering (information theory). At that time, it was common practice in linguistics to classify words not only on the basis of their meanings but also on the basis of their cooccurrence with other words. Researchers studied methods for automatically learning lexical and syntactic information from corpora, the goal being to derive an algorithmic and unbiased methodology for deducing the structure of a language. The main insight was to use distributional information, such as the environment a word can appear in, as the tool for language study. It was supposed that by clustering words and phrases based on the similarity of their distributional behavior a great deal could be learnt about language [2].

Regrettably, interest in empiricism faded in the late 1950s and early 1960s with a number of significant events including Chomsky's criticism in *Syntactic Structures* (1957) [6] and *Review of Skinner’s Verbal Behavior* (1959) [7] that redefined the goals of linguistics and initiated the development of generative linguistics.

N. Chomsky was the antagonist of statistical methods in linguistics. He argued against the learnability of language from data, believing that most of language is innate and not learned, and apparently invalidated the corpus as a source of evidence in linguistic inquiry [6]. Chomsky suggested that the corpus could never be a useful tool for the linguist, as the linguist must seek to model language competence rather than performance. *Competence* is best described as our tacit, internalized knowledge of a language. *Performance* is external evidence of language competence, and is usage on particular occasions when, crucially, factors other than our linguistic competence may affect its form. Competence both explains and characterizes a speaker's knowledge of a language. Performance, however, is a poor mirror of competence. For example, factors diverse as short term memory limitations or whether or not we have been drinking can alter how we speak on any particular occasion. This brings us to the nub of Chomsky's initial criticism: a corpus is by its very nature a collection of externalized utterances – it is performance data and is therefore a poor guide to modeling linguistic competence [5].

However, this was not the only criticism that Chomsky had of the corpus linguistics approach.

All the work of early corpus linguistics was underpinned by two fundamental, yet flawed assumptions: (1) the sentences of a natural language are finite; (2) the sentences of a natural language can be collected and enumerated.

But the number of sentences in a natural language is not merely arbitrarily large – it is potentially infinite. This is because of the sheer number of choices, both lexical and syntactic, which are made in the production of a sentence. Also, sentences can be recursive [10]. The only way to account for a grammar of a language is by description of its rules – not by enumeration of its sentences. It is the syntactic rules of a language that Chomsky considers finite. These rules in turn give rise to infinite numbers of sentences. As researcher discovered that the language is much more complex than was previously thought and as learning the complexities of language from the data began to appear hopeless, much of the work on corpus-based language learning was halted.

Chomsky’s development of generative linguistics and his critique of existing empirical approaches quickly shifted the focus away from empiricism towards rationalism, with their emphasis on symbolic grammars and innate linguistic knowledge, that is, *universal grammar* [2], in a remarkably short space of time. Mathematical or mechanistic view on language was based on the supposition that NL was a calculus described by set theory, relation algebra and mathematical logic [11]. It was expected that a language could be described by a finite set of sentence production rules and accordingly by rules of definition of generated constructions accuracy from the standpoint of their acceptance to a language system. But the problem of an acceptability of such generated constructions from the sense point of view was not considered. The algorithms based on Chomsky’s TG-grammar have appeared effective only for the analysis of restricted "subsets" of NL being inadequate for the language as a whole. It has been discovered the fundamental limitations on the description of language structure and the unsuitability of TG-grammar for the description of NL semantics [13].

Attempts to construct AI systems with the lingware on the basis of universal closed semantic languages and TG-grammar have not given positive results. As an example it can be given the destiny of widely advertised at the 1970-80s European machine translation system EUROTRA and Soviet projects ETAP-I, ETAP-II. Considering NL as "algebraic calculus", the developers hoped to receive high-performance machine translation by deepening the language formalization. In practice it has appeared that such formalization requiring the increasing human efforts did not improve, but made the final result noisy. It has resulted in crisis of dimension and gradual rejection of research [11].

Nevertheless, during the 1970s AI systems had been developed, in most cases obviously or implicitly based on generative grammars already oriented by processing of all "well-organized" sentences and only them. It’s known that at restricted data domain the semantic limitations always have a single meaning. This property was successfully used during the development of the NL-interface for databases, e.g., in system PLANES [16].

The first DARPA Speech Understanding project [9] emphasized the use of high-level constraints (e.g., syntax, semantics, and pragmatics) as a tool to disambiguate the allophonic information in the speech signal by understanding the message. Researchers hoped to use NLP techniques such as ATNs based on branch network to understand the sentences that they were trying to recognize even though the output of their front end was highly variable and ambiguous. ATNs was successfully used in linguistic processor of the first question-answering system LUNAR [17]. But the development of such systems remained labour-consuming, requiring of considerable efforts on the part of engineering of knowledge dependent on a data domain. The testing implemented on the artificially simulated linguistic situations. But there are different deviations from the standard rules in real language [13], therefore these systems could not adequately operate outside of restricted problems, for which one they were designed.

## 3. The revival of corpus linguistics

Although Chomsky's criticisms did discredit corpus linguistics, all corpus-based work had not been stopped. For example, in the field of phonetics, naturally observed data remained the dominant source of evidence with introspective judgements never making the impact they did on other areas of linguistic inquiry. Also, in the field of language acquisition the observation of naturally occurring evidence remained dominant.

In the early 1980s there was some work in automatic induction of linguistic knowledge directly from the text taking into account the statistical characteristics of lexical and grammar units of language. These activities were based largely on two widely available annotated corpora: the Brown corpus and the Lancaster-Oslo-Bergen (LOB) corpus. One of the first successes of corpus-based learning was in part-of-speech (POS) tagging, that is, assigning an appropriate lexical syntactic class (e.g., noun, verb, article) to each of the words in a sentence performed at an accuracy close to human performance (>95%) [2].

The revival of corpus linguistics closely connected with the flourishing renaissance of empiricism in computational linguistics grown out of the experience of the speech recognition community during the 1970s and 1980s. Many of the same statistical techniques (e.g., Shannon's Noisy Channel Model, n-gram models, hidden Markov models (HMMs), entropy (H), mutual information (I), Student's t-score) have appeared in one form or another, often first in speech, and then soon thereafter in language. Many of the same researchers have applied these methods to a variety of application areas ranging from language modeling for noisy channel applications (e.g., speech recognition [14], optical character recognition), and spelling correction, to part-of-speech tagging, parsing, translation, lexicography, text compression and information retrieval.

Starting in the late 1980s the success of statistical methods spread to other areas of NLP: syntactic analysis, semantic analysis, speech synthesis, information extraction [3]. Research on empirical methods is now thriving in a variety of other areas as well, such as word-sense disambiguation, prepositional phrase attachment, anaphora (e.g., pronoun) resolution, and discourse segmentation. The noisy-channel model permitting to receive the original data from noisy data has been successfully used in areas of language processing as disparate as spelling correction and machine translation [2].

With the development of tree banks, large data-bases of sentences annotated with syntactic parse trees came an increasing body of research on empirical parsing methods, e.g., probabilistic context-free grammars (PCFGs), context-free grammars in which every rule is assigned a probability. The probabilities are to be interpreted as the probability of expanding a constituent using this particular rule, as opposed to any of the other rules that could be used to expand this kind of constituent. Given the probability of individual rules, it could be calculated the probability of an entire parse by taking the product of the probabilities for each of the rules used therein [4]. As one more example it is possible to mention n-gram models which count the occurrences of word n-tuples in a training corpus. The most commonly used versions are bigram (n=2) and trigram (n=3). The joint probability of the word sequence can be expressed as a product of word probabilities each conditioned on all preceding words [15].

But empiricism spread rapidly throughout the NLP community to a large extend the result of seminal research in speech recognition. Research originating at IBM Yorktown resulted in statistical (stochastic) methods based on HMMs. These methods use a probabilistic finite-state machine [2]. Most current commercial speech-recognition systems use HMMs.

The recent revival in CL investigation has been attributed to potential solutions to several related, long-standing problems in NLP such as [2]:

- *acquisition*, identifying and coding all necessary knowledge automatically;
- *coverage*, accounting for all the phenomena in the given domain or application;
- *robustness*, accommodating real data that contain noise and different unaccounted aspects by the underlying model;
- *extensibility*, easily extending or porting a system to a new set of data or a new task or domain.

It is also necessary to mention achievements in development of computer facilities and information technologies making research realistic [8]:

- computer hardware advances which have increased processor speeds and memory capacity, while reducing prices;
- increasing availability of large-scale, language-related, on-line resources, such as dictionaries, thesauri, and ‘designer’ corpora – corpora selected for representativeness and perhaps annotated with descriptive information;
- the demand for applications in a world where electronic text has grown exponentially in volume and availability, and where electronic communications and mobility have increased the importance of multi-lingual communication;
- maturing NLP technology which is now able, for some tasks, to achieve high levels of accuracy repeatedly on real data.

Perhaps the most immediate reason for this CL renaissance is the availability of massive quantities of data: more text is available than ever before. Just ten years ago, the one-million word Brown Corpus was considered large, but even then, there were much larger corpora such as the Birmingham Corpus. Today, many locations have samples of text running into the hundreds of millions or even billions of words. Collections of this magnitude are becoming widely available, thanks to data collection efforts such as the Association for Computational Linguistics' Data Collection Initiative (ACL/DCI), the European Corpus Initiative (ECI), ICAME, the British National Corpus (BNC), the Linguistic Data Consortium (LDC), the Consortium for Lexical Research (CLR), Electronic Dictionary Research (EDR), and standardization efforts such as the Text Encoding Initiative (TEI).

The main tendency in up-to-date NL investigation is the desire to move beyond (artificially) restricted domains and systems to more realistic applications and products. Most industrial research laboratories have shifted their emphasis from basic to applied research with system evaluation and, what is more important, testing on realistic data [2]. Using the testing results experts and knowledge engineers verify the competence of systems and correct them to increase their efficiency.

Automated learning and training techniques allow much of relevant knowledge to be acquired directly from data than laboriously hand coded. If the training data are extensive and represent all the relevant phenomena, CL methods attempting to optimize performance over the complete training set, can help to ensure adequate coverage. These methods can produce a probability estimate for each analysis, thereby ranking all possible alternatives. This more flexible approach can improve robustness by accommodating noise and always allowing the selection of a preferred analysis even when the underlying model is inadequate. The CL methods allow for automatic retraining on additional data or data from a different distribution or a new domain, they can also help improve extensibility [2].

One more characteristic distinguishing CL methods concerns the type of training data required. Many systems use supervised methods and require annotated texts in which human supervisors have labeled words with part of speech or semantic sense or have annotated sentences with syntactic parses or semantic representation [1]. Other systems employ unsupervised and use unannotated text. Unsupervised learning is generally more difficult and requires some methods for acquiring feedback indirectly, such as assuming that all sentences encountered in texts are positive examples of grammatical sentences in the language.

Finally, it is important to note that a number of existing results have shown the feasibility of learning linguistic knowledge automatically from large text corpora. Empirical data enable the linguist to make objective statements, rather than those which are subjective, or based upon the individual's own internalized cognitive perception of language. CL can provide the mean to overcome the linguistic knowledge-acquisition bottleneck and make advanced language processing a reality.

## 4. Conclusion

CL approach has been widely used and is becoming popular in research communities. Its applications range from POS tagging, syntax analysis, semantic analysis, machine translation, lexicon acquisition, corpus annotation, error recovery and more. Since large training corpora are becoming more and more available, computing power can be accessed at very low cost, and statistical optimization techniques are now well developed, it is expected that the CL paradigm will be one of the most promising approaches for future natural language processing development.

## References

[1]. Boguslavski I.M., and others. “Annotated Corpus of Russian Texts: the Concept, Annotated Tools , Types of Information”, *Proc. of International Seminar Dialog’2000 on Computational Linguistics and its Applications,* **2**, (2000), 41-47. (in Russian)

[2]. Brill E., Mooney R.J., “An Overview of Empirical Natural Language Processing”, *AI magazine,* **18**, 4, (1997), 13–24.

[3]. Cardie C., “Empirical Methods in Information Extraction”, *AI magazine,* **18**, 4, (1997), 65–80.

[4]. Charniak E., “Statistical Techniques for Natural Language Parsing”, *AI magazine,* **18**, 4, (1997), 33–43.

[5]. Chomsky N., *Language and Mind,* Harcourt Brace, New York, 1968.

[6]. Chomsky N., *Syntactic Structures*, The Hague, The Netherlands: Mouton, 1957.

[7]. Chomsky N., “Review of Skinner’s Verbal Behavior”, *Language,* **35**, (1959), 26–58.

[8]. Cunningham H., Humphreys K., Gaizauskas R., Wilks Y., “Software Infrastructure for Natural Language Processing”, *Proc. of the 5th Conference on Applied Natural Language Processing,* 1997. – Available at http://xxx.lanl.gov/ps/cmp-lg/9702005.

[9]. Klatt D.H., “Review of the APRA Speech-Understanding Project”, *The Journal of the Acoustical Society of America*, **62**, 6, (1977), 1345–1372.

[10]. Mcenery T., Wilson A., *Corpus Linguistics,* Edinburgh University Press, 1996.

[11]. Piotrowski R. G., “Automatic Text Processing: the Theory and Practice by the end of XX Century”, *Scientific and Technical Information,* **2**, 5, (1998), C.26–36. (in Russian)

[12]. Piotrowski R. G., *Linguistics Engineering and Language Theory,* Leningrad, 1979. (in Russian)

[13]. Sovpel I.V., *Engineering and Linguistic Principles, Methods and Algorithms of Automatic Text Processing,* Minsk, 1991. (in Russian).

[14]. Stolcke A., “Linguistic Knowledge and Empirical Methods in Speech Recognition”, *AI magazine,* **18**, 4, (1997), 25–32.

[15]. Su Keh-Yih, Tung-Hui Chiang and Jing-Shin Chang, “An Overview of Corpus-Based Statistics-Oriented (CBSO) Techniques for Natural Language Processing”, *Intl. Journal of Computational Linguistics and Chinese Language Processing (CLCLP),* **1**, 1, (1996), 101–157.

[16]. Waltz D.L., “An English Language Question-Answering System for a Large Relational Database”, *Communications of the ACM,* **21**, 7, (1978), 526–539.

[17]. Woods W., “Progress in Natural Language Understanding: an Application to Lunar Geology”, *AFIPS Conference Proceedings,* **42**, (1973), 441–450.