PRECISION AND IMPRECISION IN MATHEMATICAL AND COMPUTATIONAL LINGUISTICS. THE EXAMPLE OF GR.C. MOISIL'S MECHANICAL GRAMMAR OF ROMANIAN

RADU GRAMATOVICI

1. INTRODUCTION

In a series of papers published in 1960, Grigore Moisil brought to our attention some of the problems encountered at the time in the field of automatic processing of natural language. In this paper, we will try to shortly investigate how the problems approached by Moisil are reflected in the current research in the (related) fields of Mathematical Linguistics, Computational Linguistics and Natural Language Processing. For a recent overview of Moisil's work and life, the reader is referred to [4].

2. THE MECHANICAL GRAMMAR OF ROMANIAN

We will discuss here four papers [5, 6, 7, 8], written by Moisil in the beginning of the sixties. The first three papers refer in principle to the automatic translation of natural languages and more specific of Romanian into another language and viceversa. The fourth paper discusses the conjunction *ŞI* ("AND") from different angles (linguistical, mathematical or using Boolean circuits).

Preliminariile traducerilor automate [5] is the first paper from the series. Mosil starts this paper by making general considerations on the way in which the written text – letters, words, phrases – are encoded in the computer. A parallel is made with the (algebraic) theory of codes.

A *dictionary* and a *grammar* for Romanian have to be written in the memory of the computer. But:

RRL, **LI**, *3–4*, p. 499–506, Bucureşti, 2006

> "E nevoie de o cercetare în spiritul maşinilor de calcul a tuturor verbelor, substantivelor şi adjectivelor româneşti, dând regulile morfologiei lor."[1]

This is what Moisil explores in the papers [5, 6, 7]. In [5], Moisil refers more to general problems of the automatic translation, while in the other two papers, he approaches specific problems, like the conjugation of verbs (in [6]) or the declination of nouns and adjectives (in [7]).

In [5], topics like the difference between the conjugations of verbs in different languages or different meanings of words in the dictionary are approached. For example, in Russian the verb in the past tense is declined according to the gender, while in Romanian or other languages, this feature is not present. Regarding the meaning of words, a word with one meaning in a language may have several meanings in other language and the computer making the automatic translation should choose the correct translation of the word, probably using the context in which the word is used.

Another characteristic of the natural language that one needs to describe is the statistical nature of many phenomena. In this respect, measures like:

- the frequency of words;
- the frequency of grammatical and stylistic forms;
- the frequency of idioms;
- the frequency of letters, digrams, trigrams, etc.
- the distribution of the length of words;

have to be studied.

In *Probleme puse de traducerea automată. Conjugarea verbelor în limba română scrisă* [6], Moisil proposes a reconsideration of the conjugation groups of Romanian verbs according to the principles of the mechanical grammar. Romanian verbs are split in five categories denoted after the vowels A, Î, E, I, U. The category A includes almost all the verbs from the 1st conjugation. The verbs from the 4th conjugation are distributed between categories Î and I. The categories E and U cover the conjugations II and III but not one-to-one distributed since some verbs from the 2nd conjugations fit in the category E, while others in the category U. As a result of this regrouping, very few verbs remain to be treated as exceptions.

A special feature introduced by Moisil in the automatic conjugation of verbs is the notion of variable letters. Variable letters are letters that may change for different forms of the same verb. The five categories in which the verbs are distributed are meant to provide a good description of the variable letters change for different forms of the verbs. The definition of variable letters is given for each category of verbs.

[1] "There is a need for a research of all Romanian verbs, nouns and adjectives, in order to describe their morphology in the spirit of computers."

We definitely may consider the variable letters introduced by Moisil as a precursor of the Two-Level Morphology, a finite-state morphological model described by Koskenniemi in 1983 [2].

In *Problèmes posés par la traduction automatique. La déclinaison en roumain écrit* [7], Moisil continues the description of the mechanical grammar of Romanian, with the declination of nouns, adjectives, infinitives playing the role of nouns, past participles used as adjectives and certain noun and adjective forms obtained from verbs and adjoined suffixes. The importance of the variable letters in the above declination is one more time restated.

In *Asupra conjuncției ŞI* [8], Moisil makes an interesting mathematical analysis of the conjunction *ŞI* ("AND"). Several axiomatic rules for the utilisation of the conjunction "AND" are given. Then, these rules are compared with formulas of mathematical logic, of algebraic calculus and of Boolean circuits, in all these framework making use of a conjunctive operation.

For example, compared to mathematical logic and algebraic calculus where the signs "&", respectively "+" may connect only two objects, in the natural language, and in Boolean circuits as well, the conjunction "AND" may connect several objects in the same construction, like in:

> "Ştefan, Ion, Maria şi Ruxandra, împreună cu fraţii, surorile şi verişorii lor, sănătoşi, veseli şi fericiţi, dansau, cîntau şi săreau ici şi colo."[2]

3. THE MATHEMATICAL PRECISION

In [5], Moisil starts to explain what the preliminaries of the automatic translation suppose. In his opinion, the first preliminary issue concerns the description of the data that has to be automatically processed.

> "Iată o primă observaţie ce avem de făcut: maşina nu poate 'înţelege' decât lucrurile precise. Maşina nu face concesii. Dacă eu scriu litera Ă maşina va scrie litera Ă şi numai prin ordine speciale maşina poate înlocui pe Ă cu A. Această '*precizie matematică*' e ceea ce trebuie să caracterizeze întreaga pregătire a traducerilor automate."[3]

In his papers, Moisil analyses some descriptive parts of Romanian, like the verb conjugations and points out that asking for mathematical precision in the description of a natural language may suppose the rewriting of the grammar of that language.

[2] "Ştefan, Ion, Maria and Ruxandra, together with their brothers, sisters and cousins, healthy, gay and happy, were dancing, singing and jumping here and there."

[3] "This is the first remark that we have to make: the machine cannot "understand" anything than precise matters. The machine makes no concessions. If I write the letter Ă, the machine shall write the letter Ă and only by special instructions the machine can replace Ă by A. This "*mathematical precision*" has to characterize the entire preparation of the automatic translation."

Mathematical precision means to define and explain all cases and exceptions found in the language, until the last detail. A statement like: "Sometimes, the subject of the phrase in Romanian can be in other cases than the nominal case." can be satisfactory in a linguistic study, even it is not accepted by all the linguists. In a mathematical description of the language, such a statement cannot be accepted. "Sometimes", in this case, has no mathematical substance. It has no computational substance either, since, in a software application, exceptional should be treated with the same importance as the regular ones.

Mathematical precision implies the completeness of the description and the lack of redundancy.

However, the mathematical characterization of the natural language is not appreciated and sometimes even not accepted by all researchers in the field, including linguists, psychologists, philosophers and even computer scientists.

In a recent paper, published in "Mind", R.T. Cook is (still) advocating the use of using formal mathematical models for describing semantics.

> "One of the main reasons for providing formal semantics for languages is that the mathematical precision afforded by such semantics allows us to study and manipulate the formalization much more easily than if we were to study the relevant natural languages directly."

The general critique brought to the mathematical description of the language is that all such attempts failed when tested on real examples or application and consequently that the mathematical precision is not necessary or even not suitable for describing natural languages.

After so many years and attempts of describing the natural language with mathematical precision, it seems clear that the main issue is not the improvement of the mathematical description, but the mathematical description itself.

The description of some data in a mathematical form supposes the entailment of the data on a mathematical structure. You cannot mathematically describe data without structuralizing it. The decision about choosing a mathematical model for describing linguistic data is crucial for the benefits that you may obtain from this description. If you choose the wrong mathematical model, than the results will be poor, regardless how accurate you describe the language.

Unfortunately, most of the mathematical models used so far for describing the natural language failed in providing a good quality of the description. Even the most successful models have a limited applicability.

Why is so difficult to choose or develop a suitable mathematical model for characterizing linguistic data?

Designing an appropriate mathematical model would imply a series of a priori decisions about the nature of the language. Usually, these decisions are determinant for the further development of the model.

Let us take the length of the phrase. This is a mathematical measure that is used for evaluating the computational complexity of parsing algorithms. When

searching for a mathematical model for parsing, one is looking for structures that have associated efficient (polynomial) algorithms in the length of the phrase. However, most of the phrases in a natural language are of a bounded length. Then, is it relevant to consider the complexity of the algorithm in the length of the phrase? What if the number of rules describing the parser is significantly greater than the length of the phrase?

Following this path, we arrive to the crucial question about the nature of the language: "Is the language finite or infinite?" Only by choosing an answer for this question and you make an important mathematical decision. If the language is finite then the problem is even more difficult, since finite languages are not efficiently implemented by current mathematical models. A solution would be the approximation (on what criteria?) of the finite languages by infinite languages (see [3]).

Probably, there is no general mathematical model suitable for every purpose. In this case, it would be more appropriate to establish a specific mathematical model for each linguistic application.

The usual methodology for choosing a mathematical model for language description is:

- A mathematical model is picked-up based on previous experience on modeling natural language;
- The mathematical model is populated with linguistic data;
- The system (data and tools) is validated against other similar systems or practical applications.

The problem issued by this methodology is that the data collection phase is very costly and time consuming as automatic data collection tools are always used in conjunction with manual processing. Are we always finishing by really validating our model on a significant data set? Obviously, no. And what happened with the huge effort spent on population our model with data, if the final result is negative (as it is the case very often)?

Our conclusion, when coming to mathematical precision in linguistics is that too much time was spent in the last 50 years with trying to improve data description on weak mathematical models, instead of strengthening the power of mathematical models themselves, before starting to populate them with data.

4. THE IMPRECISION

While mathematical precision is required for ensuring the soundness and completeness of the language description, natural languages are full of imprecise data.

In [5], Moisil refers to imprecision as a statistical property of the language.

> "Afirmația: 'cuvântul... nu aparține limbajului matematic' nu este o afirmație de același tip cu cele pe care le utilizează mașina. Am arătat că limbajul de mașină e precis. Propoziția de mai sus nu este precisă; ea nu înseamnă că e exclus ca cuvântul considerat să apară într-un text matematic (de pildă printr-o greșeală de tipar), ci că apariția lui într-un asemenea text e foarte puțin probabilă. Astfel de adevăruri sunt adevăruri statistice."[4]

The current mathematics acknowledges many types of imprecision, most of them being observed in natural languages also. Fuzzy sets or rough sets are only two of these examples of imprecise descriptions, which are not of a statistical nature. While the theory of probabilities tries to capture the randomness, fuzzy sets refer to the property of vagueness, modeling the degree of membership of an element to a set, while rough sets refer to the property of indiscernibility, modeling the (in)capacity of distinguishing between two sets of elements.

In [3], S. Marcus presents a typology of imprecision, describing much more other cases of imprecision than randomness, vagueness and indiscernibility. Abstraction, approximation, generalization, ambiguity (in a narrow sense), negligibility, plausibility, possibility, credibility, uncertainty, confidence, ignorance, absence of cohesion and lack of coherence are other forms of imprecision presented in [3].

There are opinions among researchers that once you try to describe in mathematical terms the imprecision found in natural languages, you loose exactly the character of imprecision you have tried to capture in your model. In [1], R.T. Cook quotes in this respect the position of M. Tye and R.M. Sainsbury.

> "Michael Tye and R. M. Sainsbury have argued that traditional set-theoretic semantics for vague languages are all but useless, however, since this mathematical precision eliminates the very phenomenon (vagueness) that we are trying to capture."

This critique is not meaningless, since when one gives an exact definition, one restricts the imprecision of real phenomenon to the precision of the definition.

However, as R.T. Cook explains in his paper [1], it seems that the imprecision observed in natural language is more related to the dynamics of the language construction than to the description of the language itself.

> "Here we meet this objection by viewing formalization as a process of building models, not providing descriptions. When we are constructing models, as opposed to accurate descriptions, we often include in the model extra 'machinery' of some sort in order to facilitate our manipulation of the

[4] "The statement: "the word … does not belong to the mathematical language" is not one of the statements used by the machine. We have shown that the machine language is precise. The above statement is not precise; it does not mean that there is no situation in which the given word occurs in a mathematical text (for example, by a typo), it means that the occurrence of the given word in such a text is highly improbable. Such truths are statistical truths."

model. In other words, while some parts of a model accurately represent actual aspects of the phenomenon being modelled, other parts might be merely artefacts of the particular model. With this distinction in place, the criticisms of Sainsbury and Tye are easily dealt with—the precision of the semantics is artefactual and does not represent any real precision in vague discourse."

Developing this idea, we may say that the imprecision found in the natural language is more a matter of performance, rather than one of competence. This means that imprecision reflects the incapacity of the system to reach its competencies due to some lack of information, time or resources expressed at a certain moment.

With this explanation in mind, all the formal descriptions of the imprecision reflect only the observational behavior of the language system and not the characterization of the language. The imprecision is not a cause but an effect. We observe the imprecision, we model the observation by different formalisms, but this only an artifactual characterization of the language meant to hide the imprecision of the mathematical model and not the imprecision of the language.

Consequently, we may speak about a sort of weaknesses of the mathematical models used so far with respect to the dynamics of the language. Perhaps, it would be better to look at the language as a continuous process, instead of trying to characterize it as a discrete set of objects and rules.

Considering the example given by Moisil in [5] (and presented at the beginning of this section), a pragmatic analyses of the situations in which the given word may occur in a mathematical text would probably offer a more correct explanation of the phenomenon than the simple statement that the word may occur in a mathematical text with a probability of p%. The latest is only an observation, which might be true without reflecting a certain preference of the system to the rejection of the given word from mathematical texts.

In this way we come closer to the conclusion stated at the end of the previous section: provide better mathematical models rather than trying to add artifacts to the existing ones.

5. CONCLUSIONS

Are the mathematical models necessary for the automatic processing of the natural languages? On this question one may answer with another question: is there any other way of data specification able to guarantee the completeness and correctness of the implementation than the one based on mathematical elements?

In the same time, it is worth to say that not any enumeration of cases and exceptions constitutes a mathematical model. Considering the complexity of the task, strong mathematical instruments should be employed in order to obtain decent results.

It is clear that the failure of the past mathematical models does not mean either that mathematics fails in general to describe natural languages or that there is no need for mathematical models in linguistics. The general efforts in the last 50 years were focused more on obtaining better results with rather poor (and substantially the same) mathematical instruments rather than to enhance these mathematical instruments and approach the battle with other weapons.

Of course, strong mathematical instruments are not accessible to non-mathematicians. But this is not a must.

What the computational linguistics research community lacks in this moment is a strong development framework, a kit of tools for developing efficient language model, based on strong mathematical instruments, but transparent in this respect to the users. The development of a linguistic toolkit, a suite of integrated software applications built on healthy mathematical foundations is a challenging research direction for the upcoming years.

But, probably the most important action in this respect is to bring together linguists, computer scientists and mathematicians and to put them to work in interdisciplinary teams, in which all faces of the problem are fully considered and valuated.

Or, as Moisil was saying in 1960 (see [5]):

> "Iată un vast plan de cercetare care nu se poate duce la îndeplinire decât printr-o colaborare a lingviştilor cu statisticienii, cu matematicienii şi cu tehnicienii."[5]

REFERENCES

1. Cook, R.T., 2002, *Vagueness and Mathematical Precision*, Mind, 111, 442, 225–248.
2. Koskenniemi, K., 1983, *Two-level morphology: A general computational model for word-form recognition and production*. Publication 11, University of Helsinki, Department of General Linguistics, Helsinki.
3. Marcus, S., 2004, "A Typology of Imprecision", in *Proceedings of the 1st Brainstorming Workshop on Uncertainty in Membrane Computing*, Palma de Mallorca, 169–183.
4. Marcus, S., 2006, "Grigore C. Moisil: A Life Becoming a Myth", *International Journal of Computers, Communications & Control*, 1, 73–79.
5. Moisil, G., 1960, "Preliminarile traducerilor automate", *Limba română*, IX, 1, 3–10.
6. Moisil, G., 1960, "Probleme puse de traducerea automată. Conjugarea verbelor în limba română scrisă", *Studii şi cercetări lingvistice*, XI, 1, 7–29.
7. Moisil, G., 1962, "Problèmes posés par la traduction automatique. La déclinaison en roumain écrit", *Cahiers de linguistique théorique et apliquée*, I, 123–134.
8. Moisil, G., „Asupra conjuncției ŞI", 1964, in *Omagiu Rosetti*, Editura Academiei Române, 263–266.

[5] "This is a large research plan, which cannot be accomplished without a close collaboration of the linguists with statisticians, mathematicians and technicians."