STUDIA UNIVERSITATIS

Revistă ştiinţifică a Universităţii de Stat din Moldova, 2010, nr.7(37)

A FORMAL LANGUAGE OF CORPORA

Ioachim DRUGUS

Academy of Science of Moldova

Articolul introduce un limbaj pentru reprezentarea formală a corpurilor lingvistice (plural pentru *corpus linguistics*). Acest limbaj permite reprezentarea prin intermediul conceptelor matematice a unor noţiuni specifice limbajelor formale existente şi a limbilor naturale. Este demonstrat că la modelarea limbilor naturale aparatul existent al limbajelor formale recurge, implicit, la semantică şi nu poate reprezenta aspectul pur formal al limbilor naturale. În schimb, obiectivul de modelare a aspectului pur formal al limbilor naturale poate fi atins prin limbajul corpurilor lingvistice, numit *metalingua*. Aparatul introdus poate avea aplicaţie în Webul Semantic, web de următoarea generaţie, menit să reprezinte informaţia în baze de cunoştinţe accesibile din Internet, şi care poate contribui, într-o anumită măsură, la modelarea limbilor naturale.

Formal languages were introduced by mathematicians with the objective to serve as models of the realities of three domains: the symbolisms developed by mathematical logics, the formal structures of the natural languages developed by linguists, and the artificial languages needed in information technology (IT) developed by computer professionals. Due to the initial focus on the linear space of text for representation of expressions, all notions and results related to formal languages have been formulated in terms of strings of characters and all the currently existing conceptuality, terminology and results indicate that mathematics has developed an apparatus for the linear formal languages rather than for arbitrary languages.

However, during the 20th century various frameworks for conceptual modeling like semantic networks have been designed, the linguists started using non-linear schemas of linguistic structures and the IT developed languages operating with planar diagrams or with graphics. Also, the approach called „geometrization of physics" emerged – an approach to reduce the laws of nature to the „geometry" of the World, where the geometry manifests as spatial and temporal relationships. This domain is also in need of a general language of geometric shapes.

The „theory of forms" going back to Plato, the founder of Western philosophy, places the notion of *form* in the basis of cognition. According to his theory, a 'form' refers not only to shapes in space or processes time, but also to „shapes" made up of colors, tastes, and other sensual data, as well as to ideas regarded as structures residing within mind. This theory can be regarded as the most general approach to data and, if formalized, could serve as a foundational framework for our days 'data processing' era, where processing is done not only by humans but also by artificial agents.

All these developments are indicative of the need to develop a general theory of „bodies of data", for which the term *corpora* started being widely used in linguistics of last decade. This paper develops a framework for corpora and is a continuation of the previous several publications [1-5]. This framework consists of a conceptuality and a formal language of corpora called *metalingua*. No previous knowledge of the material in mentioned publications is required, since this paper introduces all the required concepts.

Basic notions and terminology

Entity is the most general term used to refer to anything. An *agent* is an entity – human, animal or mechanism – capable of action. *Reflection* is a partial manifestation of action – reflection is representation of entities in mind of a human or animal or in a system which plays the role of mind in a mechanism called agent. We will say this to be a *presentation system*. Mind is a presentation system, which serves as residence for *ideas* which in turn, according to Greek etymology „idein", „to see", denote the images the mind can „see" and which, according Latin, should be called *forms*. Thus, ideas are the inner forms which represent external forms. For communication of forms other presentation systems can be used (different from mind), and among these, most frequently used are the *linear* presentation systems – written text, speech or electronic transmission of characters. An entity is represented in mind by an idea said to be „idea of the entity". We will consider correct the expression „idea of" and not the expression „idea about", since an idea is the image of an entity at a representation, and not a discourse about that entity.

An agent can discuss about entities of a class without having an idea of any particular entity. If he has an idea of a particular entity, this entity is usually said to be *object*. An object may change over time and still be considered to remain „the same”. The „sameness” of a changing object is provided by one entity in mind, or an unchanging part of the object, which we will say to be *identity* of the object. Identity and identification are matters of representation (not of knowledge, as this is discussed below). For a rigorous approach, we will consider that an object is an ordered pair (*identity*, *content*) and the *identity* of the object *identifies* the *content* of the object.

To have an identity of an object does not mean to be always able to identify the object. Say, as long as we watch a star, no matter that it changes its luminosity, we recognize it as the same object (here, the identity is actually provided by continuity of observation). But the next night, we might not be able to locate it in sky, and even if we have been indicated the right star and we kept in mind its former identity, we might not be able to confirm that it is the same star as the one watched previously. In order to recognize the star, we must also have knowledge about how to recognize it. Such knowledge consists of the star luminosity and relative position with respect to other stars of the given star.

Knowledge is more than *idea* – knowledge about an entity *e* is composed of ideas and is about *e* (unlike the „idea” which „of”). A notion is an entity *n* whereby an agent „knows” another entity *e* (see Latin „notion” from „noscere”, „to know”). Knowledge is composed of notions and is a matter of cognition of the content (not of representation like an identity). *Recognition* is possible due to knowledge and not only due to representation. The entities which are exposed to the process of recognition are the *objects*, i.e. entities for which an agent already has identities.

We adopt the following terminology related to recognition: a set of properties owned exactly by the elements of a subclass *D* of a class *C* is called *pattern* of the elements of a subclass *D* in the class *C*, or just *D-pattern* in *C*. An identification pattern is a pattern of any one-element subclass of *C*. We will say an object to be *identifiable* by an agent if the agent, additionally to having an identify for the object, also has an *identification pattern* for the object.

An object *varies* and this is why the problem of identification comes up. We will say that in variation the object gets various *forms*. Since identity is a matter of representation, we cannot say that it is the identity which varies – we will say that the content of the object varies. Generally, we will refer to ‘variation according a pattern’ by *paradigm*. There are different types of paradigms. To refer to the particular type of variation of an object which relates with object identification we will use the term *identification paradigm*.

Bodies and Corpora, Symbols

We will say a *body* to be an object in space. The manner how space affects the object in order that the object is said to be *body*, is the *organization* – a space induces certain organization of the objects and their components. „Corpus linguistics” is the currently used term to refer to a body of texts. We will consider a *corpus* to be a body represented as an expression of a language – thus, we correlate the notion of *body* with the notion of *language* in order to obtain the notion of *corpus*. The expressions of a language must be represented in a presentation system in order to exchange them with other agents or to save them for the same agent to retrieve them later. Both cases have to do with *communication* – either between different agents or of the same agent with himself.

We can consider physical bodies expressions in a language of attributes of bodies, like location in space, time or scale, colors in visible or invisible spectrum, etc. Physicists limit themselves to a very small number of such attributes expressed in numbers and called *magnitudes*. Painters have palettes of colors, musicians have musical notes and wine tasters have their complex gustative denotation systems. From these examples, it becomes clear that what we call „bodies” can be treated as expressions in a language and such expressions can be said to be *corpora*.

What is a language? We will say a language to be a system of symbols or a *symbolic system*. Natural languages with their alphabets, grammars, and other linguistic schemes, as well as palettes, musical notes systems or gustative qualities denotations, and conceptual systems of logicians are symbolic systems and, therefore, are languages. All these languages are related to various realities – any reality, imposes a certain symbolic system and can be treated as a language. We will treat the notion of *symbol* as orthogonal to the notion of language. In other words, we will consider that a symbol can be present in different languages.

Symbols are complex phenomena, and we will focus here only on several aspects, which are important for this paper. „Symbol” and „sign” are generally treated as synonyms. „Sign” is linguistically related to

„semantics" and is preferred by some authors. But „sign" has been extensively used to refer to symbols without own value and used only for modifying the value of other symbols. Say, „signs" are said to be the characters "+", "-", which modify a value of a number. Therefore, I will prefer here the word "symbol".

We will say a *symbol* to be an object used in communication, which represents („symbolizes") one or several things of a class. We will say that such things make up a class, and this class is associated with the symbol. We cannot say such class to be the „class of the symbol", because this would imply that we attribute the symbol itself to a class, while we actually associate a class of (other) entities with this symbol. We will say the class associated with a symbol to be *concept* of the symbol. To have knowledge of a symbol is to have two ideas (see above about what is „to have an idea of something"), namely: 1) an idea of the „symbol itself", and 2) an idea about the concept of the symbol.

Firstly, we need to find out what might be the idea of the „symbol itself"? As any object, a symbol must have an identity. But unlike other objects, a symbol is an object destined to be used in communication and, thus, it must be *identifiable* and *reproducible*. For an object to be identifiable, the agent must have an identification pattern. As to reproduction of an object, the agent can exploit variation of form in order to encrypt various hues of sense. Thus, the variation pattern of a symbol can also have a classification of forms. The forms can be classified per natural languages and the class of forms used in a natural language to represent a symbol will be said to be a *lexeme*.

Secondly, we need to find out what might be the idea an agent must have about the class associated with a symbol said to be „*concept* of the symbol" in order to have knowledge of the symbol. A symbol is an object, i.e. it must have an *identity* and *content*. The content of a symbol is a concept which represents a class, and we will refer to the elements of this class (or „instances of the concept") by *values*.

To refer to the variation of values of a symbol according a pattern, we will use the term *conceptual paradigm*. How a concept varies is a complex phenomenon and this concept does not lie within the focus of this paper. The notion of *conceptual paradigm* was mentioned here just to give an idea about the complete inventory of aspects of a symbol.

We will prefer the word „expression" to refer to corpora when these are regarded in the context of a language. „Expression" is the commonly used word to denote something which „expresses a meaning". To justify the usage of the term expression the way we did (i.e. formally, without an associated meaning), we can also say that an expression „expresses a form", i.e. by its aspect it shows how it was built out of other expressions.

Having chosen the term „expression" to replace the term „corpus", the question arises why do we need at all the term „corpus"? Actually, the term „expression" is used by in this paper for theoretical considerations, but in practice, we encounter situations, when we need to emphasize that (1) a corpus is very large, like a collection of texts, and we need to oppose it to its smaller constituents, or that (2) a corpus contains „factual" data, which serves as „test data" for checking adequacy of a linguistic schema or for checking correct functioning of the software.

Languages and their presentations. Metalingua

A language is said to be a symbolic system and a symbol is a compound notion, where all the components are abstract. The components of a symbol, which are the closest to being concrete, is the lexeme, which is a class of forms. But even a form is an abstraction - concrete are only the manifestations of a form in a presentation system. We will say such manifestations to be *inscriptions* in presentation system. Thus, we treat the objects of a language as abstract and the inscriptions as concrete. The corpora are abstract entities and they reside in an abstract presentation system - the mind. The inscriptions of corpora are concrete entities and they reside in a presentation system.

The expressions of natural languages are strings and they are usually inscribed in the linear presentation system of text. The corpora of the UML (Universal Modeling Language) called diagrams are usually represented in a planar (2 dimensional) presentation system. But the expressions of English can be written in various manners in cross-words, in circles, or other sophisticated modes in various other word games, and they remain expressions of English. If a new practice of English is promulgated which requires to write expressions from right to left, we cannot say that English changed – what changed is the *presentation* of expressions in English. Also the UML diagrams can be represented in linear notations and these notations cannot be said to be expressions of a language different from UML. In spirit of these examples, we will separate a language from its presentation. Our treatment of expressions as *abstractions* and their inscriptions in presentation systems as *concrete* manifestations, contributes to this separation.

We will consider corpora residing in an abstract presentation system (for human, this is mind), and consider the corpora and their components to be organized by an abstract space. What is the organization of an abstract space? According our vision, the abstract space organization imposes to treat

(1) two entities a and b (presented in this order) as making up an ordered pair, which we will denote by $(a : b)$;

(2) a set of entities $a_1,..., a_n$ as totally unordered and making up a set $\{a_1,..., a_n\}$;

(3) a set of entities or an ordered pair W of entities as a *unity*, i.e. as the last entity in process of building „a unit”, and denote such unit by $[W]$.

We will consider each of the 3 types of organization types as obtained by application of an operation. These operations are called *association*, *aggregation* and *atomification* and are denoted, respectively, by parentheses with a colon as the separator of its two arguments, braces with comma as a separator of an arbitrary number of arguments, and square brackets which enclose a unique argument.

In some situations, it is convenient to omit the colon within the notation of an association, and in all situations is convenient to omit the parentheses and braces enclosed by square brackets, namely, to write $[a : b]$ instead of $[(a : b)]$, and to write $[a_1,..., a_n]$ instead of $[\{a_1,..., a_n\}]$;

The three operations are called A3 operations [1] and the language of the three types of brackets is called *metalingua* [5]. In practice, a corpus can be built by using operations different from the A3 operations but, according the vision substantiated in [4], any other operation can be represented as a superposition of the A3 operations. A corpus might be very complex in representation, say, because it requires a multi-dimensional representation system. But metalingua requires a linear representation system, the inscriptions of its expressions are linear, and thus it allows notation of any corpora of any language in regular texts. This places metalingua in the relation of „meta” to other languages and justifies its name of being „about” other languages.

Linear formal languages

Mathematics offered a formalization of languages called *formal languages*. A formal language is a set V called *vocabulary* with its elements called *vocabulas*, and a set of strings of elements of V called *expressions*. The terminology varies with different authors, and here I have chosen to use one set of terms which looked most appropriate. With formal languages, mathematics did not separate the notion of language from the notion of language presentation and considers the linear order of text as part of the apparatus of such formal languages. Thus, the formal languages, as they are generally treated in literature, are intrinsically *linear*.

The linearity of current formal languages is responsible for the ambiguity of the syntactic analysis of their expressions. A general method of disambiguation is to indicate the association structure of an expression – such a structure is convenient to be denoted in metalingua with omitting the colon. For example, the formal expression *abra* can be disambiguated by one of the following association structures: $(a(b(ra)))$, $(a((br)a))$, $((ab)(ra))$, $(((ab)r)a)$. The LISP programming language for AI imposes by default the first disambiguation, which is called *right association*.

The linear formal languages employ the linear organization of the presentation system and, thus, economize the notation. So, a linear expression can be obtained from one of its association structures by omitting the parentheses. The original expression can be restored if a default, like left association, is established. The economy is a strong argument in favor of linear formal languages.

But the linear formal languages, even with a default regarding association, also impose an organization or an expression which might not be meant by the author of the expression. For example, the corpus $\{a, b, c\}$ can be represented 6 different manners (namely, *abc, acb, bac, bca, cab, cba*), each imposing its organization different from others and different from the non-order organization of the corpus. In natural languages an aggregation like the one in the example above lies in the so called *deep structure* of the language. Thus, in representation of syntax of a natural language via formal languages, the aggregation can be discovered only by recurring to semantics.

Due to linearity, formal languages use only one operation over two expressions *e* and *f* – the so called *concatenation*, which is denoted by *e* * *f*. This operation is *associative*, i.e. for any *e, f, g,* the statement ((*e* * *f*) * *g*) = (*e* * (*f* * *g*)) is true, and this allows to drop parentheses, which is responsible for the ambiguity or syntactic analysis.

The linear formal languages have only two means to indicate a *unit* which in corpora is obtained by the operation of atomification. Namely, something is a unit only if it is either member of the vocabulary or if it is an expression - a formal language is essentially a language with two sorts of objects. To work with expressions made up of other expressions, one has either to work with two formal languages instead of one, or to add to the vocabulary of the formal language a separator, which separates „lower level" expressions to produce „higher level" expressions. As an example of the last phenomenon can serve the phrase: *rev up the engine*. Either which two formal languages, or with a formal language enhanced with a separator, in both cases there is no means to show whether *rev up* should be treated as the atomification [*rev up*] or as the association (*rev up*). An English speaking person knows, that the expression *rev up* should be treated as a unit (atomification) with the meaning *to accelerate* and the word *rev* cannot be followed by any other word.

Conclusion

The formal languages as defined today, and designed for mathematical purposes, are linear, are limited in their expressive power, and impose unavoidable ambiguity. Such drawbacks can be overcome only by the semantic analysis. Therefore, by means of currently existing formal languages the syntax of natural languages cannot be completely separated from semantics.

The proposed formal language of corpora said to be *metalingua* manages to model the syntax of natural languages without recurring to their semantics, and to separate the surface structure from the deep structure of a language. Therefore, metalingua is a better apparatus for modeling natural languages syntax than currently existing formal languages.

References:

1. Drugus I. A Wholebrain approach to the Web // Proceedings of "Web Intelligence and Intelligent Agents International Conference", Section „New Computing Paradigms for Web Intelligence and Brain Informatics", Silicon-Valley, November 2-5, 2007, p.68-72.
2. Drugus I. Universics – a structural framework for knowledge representation // Proceedings of „Knowledge Engineering: Principles and Techniques", International Conference, Cluj-Napoca, Romania, July 2-4, 2009, p.115-118.
3. Drugus I. Universics: an Approach to Knowledge based on Set theory // Proceedings of "Knowledge Engineering Principles and Techniques, International Conference, Selected Extended Papers", Cluj-Napoca, Romania, July 2-4, 2009, p.193-200.
4. Drugus I. Universics – a Common Framework for Brain Informatics and Semantic Web // Chapter 1 in the book "Web Intelligence and Intelligent Agents", Ed. InTechWeb, 2009, p.1-24.
5. Drugus I. Metalingua – a Formal Language for Integration of Disciplines via their universes of discourse // E.T.C. Journal, 2009, p.17-23.

Prezentat la 23.03.2010