Studies in Communication Sciences 1 (2001) 225-238

STEFANO CRESPI-REGHIZZI *

FORMAL LANGUAGES IN COMMUNICATION SCIENCES: A MANIFESTO

Based on actual experience at USI, it is argued that a course on formal languages should play a central role within the university curriculum in Communication Sciences. The course is valuable not just for information technology tracks, but especially for humanities: linguistics, semiotics and human communication. Therefore this course is very different from formal language courses taught in computer science/engineering curricula, that had provided initial inspiration. The topics covered are outlined and related with the skills students are expected to acquire. Last the view is expressed that formal language theory jointly with logic should be preferred to traditional mathematics as a mathematical basis for students in communication.

Keywords: formal languages, formal theories for communication sciences, mathematics for communication studies, linguistic abstraction.

Technological and mathematical courses in communication sciences

New curricula in communication sciences have been recently started by several universities in response to a strong request for education determined by the diffusion of multimedia networks.

* Stefano Crespi-Reghizzi, Politecnico di Milano (Italy); Faculty of communication sciences, University of Lugano (Switzerland), email: stefano.crespi.reghizzi@lu.unisi.ch.

The curricula should prepare the specialists of communication involving public institutions, companies, and persons in their roles of citizens, voters, tax-payers, customers, information providers or, in general, users of services.

At USI the curriculum in Scienze della Comunicazione, one of the first established three years ago, offers a peculiar, rich blend of humanities and technologies. Side by side with linguistic and semiotic and with sociological, historical and psychological disciplines, one finds a significant presence of information technologies (IT), namely informatics and telecommunications.

We recall the three existing tracks of the USI degree: Mass communication and new media, Enterprise and institutional communication, Communication technologies. Without getting into details (see [1]), notice that there is a nucleus of basic technological courses that is required by all tracks, whereas other courses are compulsory for technology students, and a third group is optional for the first or second track.

In the original curriculum (this writer does not deserve credit for that) it is not difficult to recognise that IT courses derive from the classical disciplines one finds in Computer science/engineering curricula. Most titles are the same: in particular Electronics, Databases, Knowledge engineering, Software engineering, Operating systems, and last Formal languages, of concern here. But in reality the programs of the courses at USI are not the traditional ones of computer science departments, because the educational objectives are different. In fact the careers of future communication professionals will be distinct, yet partly interchangeable with the careers of IT graduates. Currently the main perceived vocation of USI graduates is to plan and manage communication and to act as mediators between computer professionals and communication officers.

As it happens, the programs of technological courses have undergone change in the past two years. Our recent teaching experience has influenced the reshaping the curriculum, shifting the focus of "Formal languages" from technical contents and objectives to more conceptual and foundational ones. We are going to argue that this new orientation is central for communication studies.

Formal languages for communication sciences

It is well known that the theory of formal languages has emerged as a new scientific discipline in the years 1950s, from the confluence of the hitherto separated studies of linguistic, mathematical logic, and informatics (ante litteram). Emblematically, we may name as the founders: the linguist N. Chomsky, the logician S. Kleene, and the computer scientists who designed the first programming languages (J. Backus among others). Since its detachment from linguistics and logic fifty years ago, the discipline has become a pillar of IT. It has been instrumental in developing, with increasing rigour and effectiveness, countless artificial languages with so many purposes: programming, databases, artificial intelligence, knowledge representation, documents and hypertexts, robotics, communication protocols, iconic, graphic and pictorial languages, etc. In all universities computer science curricula include a course on formal languages and automata that teaches the principles and techniques needed to design artificial languages and their translators (compilers). In most universities, the course on formal languages and automata theory is preceded by introductory units on programming (and of course by mathematics), and is followed by a more practical course on compiler design.

Evidently, our framework is altogether different since on one hand communication science students lack the deep knowledge of electronics and mathematics of computer science students, on the other hand our graduates will unlikely take the specialist jobs as designers of language processors (compilers, interpreters, etc.). What now seems evident took two years to be clearly understood. Accordingly we consider as reductive the initial orientation: placed about midway in the four years curriculum for the technological track only, the course was a simplified and mildly redirected version of a classical course for a computer engineering degree (Crespi-Reghizzi:1990; 1996; 1997). Truly we had immediately realised that the artificial languages of interest for our objectives are not Java, C, or Pascal, but the languages like HTML and XML used to represent hypertexts and web documents.

But we had missed the fact that in our studies this discipline should not be engineering oriented, to design languages and compilers. Its primary value is for describing and comparing information and communication processes with rigour and accuracy: or if you wish, the goal is to acquire a conceptual tool for reasoning, not a technique for designing computer programs. Our student should acquire the ability to recast communication phenomena into a rigorous and simple model, rather than learn the algorithmic aspects of syntax and semantic analysis. From this, several important indications can be drawn.

- The course on formal languages should be offered to all tracks: Mass communication and new media, Enterprise and institutional communication, and of course Communication technologies.
- The mathematical nucleus of formal language and automata theory (together with logic) should be used to provide a minimal but appropriate mathematical education to all students.
- The prerequisites (mathematics and programming) for taking the formal language course should be kept to a minimum, which allows to anticipate the course in the first year, thus providing all students with early capacity for formalisation.

The above guidelines are under experimentation and the results will be monitored and analysed.

Contents and objectives

Roughly the course presents two focal points:
- formalisation of language;
- language transformation and processing.

In both cases "language" is not necessarily a human language, which actually is the hardest or altogether impossible to be formalised. Rather by language it is meant any message, document, text or hypertext, made by atomic symbols.

Other promising but less consolidated viewpoints will be touched later on.

Formalisation of language

The syntactic methods accurately specify the valid phrases of a language with varying degrees of precision. The concepts of string (a sequence of symbols), of formal language, and of generative grammar enable the student to model familiar situations as diverse as the structure of a web page or the temporal series of operations of an automatic cash teller. The attentive student will be able to compare the formal methods with the more descriptive ones used in other courses, such as linguistics, or theory of argumentation with its grammatical models [2]. The mathematical prerequisites are minimal: some set theory and propositional logic (Boolean algebra). The methods to be taught are regular expressions, finite-state automata, and Chomsky's context-free (also known as Backus Normal Form) grammars. But we would like to consider other approaches to formal languages (e.g. categorial grammars or associative language descriptions) that might be more appropriate to communication systems, though less popular in computer science.

Our experience has shown that most students appreciate the awareness of becoming able to model various situations that had previously escaped their describing capability. As a bonus, the students develop the ability through formalisation exercises to express in Italian (or another language) the rules of well-formedness of a document.

The hierarchical decomposition of grammars teach them the *divide et impera* strategy, the art to decompose a complex phenomenon into simpler sub-problems. This is the very base of all sound design methods and is expected to prepare the students for the system design methods of later courses in the technology track.

Reasoning by induction is also a useful exercise, as when the student is requested to infer the general rules of a language from a few examples and counterexamples.

Moreover formal language theory revisits the important notions of structure learnt in the course of General linguistics: ambiguity, structural adequacy. The concern for structural properties calls the student attention to the misunderstandings that occur when the source and destination of a message are inconsistently or deceivingly formalised. Such concepts could be also related, though less directly, with the study of coding (in the sense of Shannon's theory of information [3]), known to the students from the course of "Information systems and technologies". Studying structural adequacy, the student realises that the simplest definition is not always the suitable one, and that the complication of a document has to be paid by the complexity of the formal model.

Last, this theory is very elegant and always surprises newcomers, as it surprised the scientific world fifty years ago by the discovery of the equivalence of three altogether unrelated models: Kleene's regular expressions, Chomsky's type 3 grammars and finite automata. This brilliant mathematical result can be appreciated with very little mathematical apparatus, and somehow mitigates the mathematical prejudice of many students of humanities.

Language transformation and processing

Through the observation of several concrete languages (e.g. phone book, HTML documents, composition of a railway train, steps of non-verbal behaviour) the student realises that, disregarding surface representations, the essential syntax structures are very, very few: lists, hierarchical lists, nested or parenthesised structures, and a few others. The bewildering and confusing variety of languages is only apparent or superficial, an epiphenomenon of the "syntax sugar" that hides the deep structure. The discovery of essential properties by removing details is called language abstraction; it is a powerful conceptual instrument in many disciplines. First it brings conceptual economy into the study of languages, perhaps as Latin facilitates the study of neo-Latin languages. In class language abstraction is illustrated by several actual cases; one is the language of functional terms, to be found in many languages such as Java, predicate calculus, database query languages, and knowledge representation languages. Functional terms can be represented using parenthesised expressions, with infix operators, or with parenthesis-free Polish representations. All representations preserve the structure but differ with respect to ambiguity and conciseness.

Last but not least the relation between forms and contents, syntax and semantics to be found in any communication phenomenon, stems from all this. Syntax transformations then permit to filter or to bring to evidence the various readings of the same document. The meaning of a phrase is viewed as its translation into a second language, obtained by means of a syntax-directed transformation. The study of direct and inverse transformations highlights the properties that allow the original message to be reconstructed. Mathematically, the theory of translation provides nice examples of the concepts of mapping: direct, inverse, one-to-many, etc.

On the technological side, notice that language abstraction is used in Web documents, to separate contents from representation; an approach practically experienced by first year students of "Fundamentals of informatics". The web languages HTML and XML provide an important, practical illustration of the distinction between language and meta-language, which is often considered by the linguists as a speculative philosophical concept of interest to logicians only.

On the practical side, the students understand the techniques used to implement compilers, parsers and the software tools used to transform documents. Understanding is reinforced by participation to a laboratory session on the automatic construction of a parser for a toy language, using a parser generator. But one semester in the first year is too short and too early to acquire design skill, except by some especially motivated and software proficient individuals. Nonetheless the concepts of deterministic transformation and algorithmic complexity are introduced. In the future, should language technology become important for communication related professions, it would be easy to introduce a second advanced course in the curriculum. It is not a dream to imagine that graduates who are able in organising communication will be highly qualified for designing special-purpose languages.

Logic as a formal language

Last year the IT faculty members reconsidered the curricular requirements in mathematics and formal theories[1]. One decision was to include some mathematical logic inside "Formal languages". We all know that logic has a fundamental place in

[1] Previously technology track students were required to take a course in Mathematics borrowed from the School of Economics. The other two tracks had no mathematics. The only other course of mathematical nature is Probability and statistics.

computer science, since computer is a logical machine and programs are executable logical formulas. But it would take some time to analyse its role for our studies too. Brevity compels us just to mention some courses where references to logic are more perspicuous: "General linguistics", "Theory of argumentation", "Theory of mass communication", "Text typologies and expression techniques", and of course the IT courses, "Knowledge engineering", "Software engineering" and "Databases".

Logic is undoubtedly the oldest attempt in language formalisation, partial as it is; it concerns the statements (or propositions) having the property of being true or false. It is not difficult to present propositional calculus and first-order predicate logic as two examples of a language defined by means of a formal syntax. The truth value of a phrase is then the meaning computed by means of a syntax-directed translation incorporating the truth tables of logical connectives such as *or* and *not*. This is enough for showing how to translate back and forth from Italian to logic. But it is insufficient to cover logical reasoning, more difficult topics, which requires a rather complex formal apparatus to explain axioms, inference rules and some proof theory. This part of logic is included in "Knowledge engineering" a third year course for the technology track. The choice of topics and of presentation is still partly open: certainly it will have to be different from the courses taught in the departments of mathematics, computer science, and philosophy. For instance the use of axioms and rules of inference as classically done to formalise reasoning does not seem to fit our didactic aims. A more attractive approach is to use semantic tableaux and *reductio ad absurdum.*

Last a common, practical problem: the lack of textbooks suitable to our objectives and students, both in breadth and depth.

We do not know of existing books on formal languages[2] or on mathematical logic[3] that could be adopted, not only because they are too difficult or technical, but also because they do not present the topics and examples that would motivate students of communication.

Non-textual languages too?

The recognised importance of visual and multi-medial communication is witnessed by the presence of several courses in the USI curriculum. We wondered whether some time should be spent to present formal models for describing non-textual languages, of which various categories exist: iconic, pictorial, graphical languages of two and three dimensions; and also man-machine interactions or codified non-verbal behaviour. The idea is quite attractive, also because it is related with the other courses having to do with such forms of communication: "Visual communication", "Theory and technique of new media", "Analysis of medial and multi-medial messages". Browsing through their bibliographies we found references to models inspired by formal language theory, generative grammars and automata. But considering that formal grammars for describing non-textual languages are rather complex and less consolidated

---

[2] The classical books on formal languages such as J. Hopcroft and J. Ulmann, *Formal languages and their relation to automata*, Addison-Wesley 1969, or A. Salomaa, *Formal languages*, Academic Press 1973, are intended for mathematically proficient students and do not include illustrative examples. On the other hand books on syntax, semantics and compilation (like A. Aho, R. Sethi, and J. Ulmann, *Compilers, principles, techniques, tools*, Addison-Wesley 1986) are intended for programmers.

[3] Books on logic are, roughly speaking, of four kinds, none suitable to our environment: formal treaties, introductory material for casual readers, philosophical books, and textbooks for computer sciences.

than grammars for textual languages, we decided not to cover them. Putting too much heterogeneous material in a first course would jeopardise the understanding of essential concepts, at the risk of becoming a superficial survey of theories.

But the debate remains open and the final decision to include or not visual languages will have to take into account the relative importance of visual communication in the vocation of our graduates.

A mathematical nucleus proper for communication sciences

We quote from the already mentioned discussion, promoted by Luigi Dadda, head of the Institute of communication technologies, on mathematics in our studies:

> "The theoretical base (in the meaning of formal theories) can be viewed with respect to communication in general and especially communication by means of IT (information technology). In fact there is a growing trend for mediating the interactions between individuals or organisations through informatics and telematics. Of course theoretical knowledge is not enough for an effective use of IT, as specific operational capabilities are needed: but this is not the subject of this discussion. Good theories constitute the mental tool for *dominating the complexity* of IT systems, and for reducing the learning stress imposed by ever changing technology.
> In such a dynamic paradigm, theoretical models too undergo changes and extensions: just think of the recent development of the theory of computational complexity with its fundamental applications to communication security. In a university program it is nevertheless convenient to choose the more classical theoretical concepts, which have been consolidated by experience, without forgetting to regularly verify their actuality and adequacy [...] In order to gain in consistence and acceptance after the promising starting phase, the studies of communication science should improve the systematic and repeatable nature of the mathematical educational offered to the students".

What is the mathematical and formal knowledge that would be profitable to our student[4]? Let us name the mathematical subjects proposed by other university curricula, such as business economy, biology, and computer science (excluding mathematics itself). We find linear algebra (analytical geometry), analysis (differential and integral calculus), probability theory and statistics, and perhaps numerical calculus. In computer science we find in addition: abstract (or modern) algebra, mathematical logic, discrete mathematics. Consider the first list, often named classical mathematics: some subjects (e.g. numerical calculus) are clearly irrelevant for our curriculum. For other subjects, elementary knowledge is sufficient, at the level that is (or should be) available from high-school education. This is particular true of the mathematical apparatus needed to describe physical systems. Similarly mathematical analysis, as taught in economy and business schools, is less relevant for our curriculum, although some concepts are needed for understanding economical models. It remains probability theory and statistics, which has an important use for all tracks, and is present in our curriculum. For classical mathematics a short start-up course was deemed sufficient to recall concepts learnt in school, also to make the student base line more homogeneous. Teachers of advanced courses have been charged to introduce any mathematical concepts and notation they might need.

The second list is closely related to the present proposal, since formal language theory can be viewed as a branch of discrete mathematics, which is founded by, and supports algebra and logic.

---

[4] We refer primarily to non-technological students. Singular technology courses would require more mathematics, an issue out of the scope of this paper.

This having been said, I hope the reader will take the next statement with the indulgence due to passionate love for a scientific discipline.

Formal language theory is the formal theoretical basis of communication studies, in the same way that, say, mechanical engineering is supported by the mathematical models of physics. The difference is that communication phenomena can be formalised to a much more limited extent, because they involve human minds and complex psychological behaviours.

As all mathematics does, formal language theory (jointly with logic) reinforces the attitude to precise reasoning. The main conceptual skills that can be practised have to do with the following items:

- recursive and iterative composition
- declarative definition and algorithmic ones
- minimal and derived operators
- closure properties of infinite sets, partial orders
- ambiguity and indeterminism
- meta-definitions
- recognition of structural analogies
- mappings, functions and transformations
- problem-solving based on problem decomposition and inductive inference.

All that can be done in the simple framework of formal languages without getting confused by irrelevant details. But more than one semester would be needed to effectively communicate this rather lengthy list of skills.

A remark: the limits of logic reasoning and formalisation should be clearly enunciated in order to make students well aware of the difference between thought and communication, natural language and codified artificial languages [4], thus avoiding the pitfall of technological idolatry.

If our short teaching experience at USI has some value, we are ready to bet that humanities-oriented students too, though in general poorly motivated by mathematics, will take interested in a subject, though a mathematical one, having to do with words and texts, linguistic structures and meaning.

In conclusion it is hoped that this writing will contribute to promote a debate on the theoretical and formal bases of communication studies.

References

S. CRESPI-REGHIZZI (1997). Linguaggi formali e artificiali: aspetti sintattici, Milano: UTET.

S. CRESPI-REGHIZZI (1996). Le grammatiche ad attributi: semantica dei linguaggi artificiali, Milano: UTET.

S. CRESPI-REGHIZZI (1990). Compilatori, interpreti e tecniche di compilazione, Milano: Masson.

GRIZE J. B. (1982). De la logique à l'argumentation, Droz.

LO CASCIO V. (1991). Grammatica dell'argomentare, Firenze: Nuova Italia.

PIERCE J. R. (1983). Teoria dell'informazione, Milano: Mondadori.