Natural Language Processing 10

Modelling Language

Sylviane Cardey

JOHN BENJAMINS PUBLISHING COMPANY

Modelling Language

Sylviane Cardey
Institut universitaire de France and Université de Franche-Comté

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Cardey, Sylviane.
Modelling language / Sylviane Cardey.
p. cm. (Natural Language Processing, ISSN 1567-8202 ; v. 10)
Includes bibliographical references and index.
1. Communication models. 2. Semiotics. 3. Interlanguage (Language learning)
4. Computational linguistics. 5. System theory. 6. Natural language processing (Computer science) I. Title.
P99.4.M63C37 2013
006.35--dc23 2013000369
ISBN 978 90 272 4996 8 (Hb ; alk. paper)
ISBN 978 90 272 7208 9 (Eb)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Preface

The model presented in this book represents many years of research and direction of research.

I wish to thank Yves Gentilhomme who taught me micro-systemics. I decided to use the notion of micro-system for my research and build micro-systemic linguistics, which I based on logic, sets, partitions and relations, as a theory to apply to language analysis and generation. Realising that representing languages globally was a Utopia, but a nice dream, I thought that starting from what we wanted to demonstrate or solve might be a better way forward.

What does starting from a goal mean? It means finding the necessary phenomena and elements needed to solve the given problem. We do not need to refer individually, for example, to all syntactic phenomena to make the agreement of the past participle in French, but only to the part of syntax to which the problem is related, and likewise only the part of morphology and lexis, including semantics, concerned with the resolution of this problem.

The theory also proposes working in intension instead of extension which means working at a high level of abstraction, partitioning the elements of the language according to where they belong, whether syntax, morphology or lexis, but more often to all these phenomena at the same time. Having gathered the data and classified it, a micro-algorithmic system will then define the processing in such a way that not only will it lead to the problem's resolution, or represent what we would like to demonstrate, but it will also give us traceability which is in any case needed for scientific justification but mandatory for safety/security critical domains. A good example is our Classificatim sense mining system.

To avoid the usual question concerning a theory "Could you give an example?" the book is indeed furnished with many examples. As a result of the Erasmus Mundus programme, I have had the great fortune to have in my courses excellent students, selected after a severe admissions process, of very many nationalities, and who were curious as all good scientists are.

During my academic career I have supervised over forty PhD students, and continue to do so. These students have come from all over the world, Asia, Europe including Russia, North, Central and South America and the Middle East, and with backgrounds predominantly in linguistics, mathematics and computer science. Research is never accomplished alone, and I am indebted to them for the

majority of examples concerning the many and diverse languages that they have analysed using the theory during their research. They will recognise their contributions in the book. I wish to acknowledge my gratitude to all my colleagues who have acted as external PhD examiners, and also those who have participated in several major projects based on the theory. A great thanks too to all my colleagues who have organised and chaired the XTAL conferences that I inaugurated hoping to bring together linguists, mathematicians and computer scientists who often have difficulty in listening to each other. The conferences have taken place in France, Italy, Portugal, Spain, Finland, Sweden, Iceland and Japan. I thank too my industrial colleagues, Airbus Operations SAS, who have aided me in understanding what are real applications in safety/security critical domains. Words cannot say how much I am indebted to my husband, Peter Greenfield, colleague and companion in research.

I would like to thank Professor Ruslan Mitkov, the editor of this series, along with the anonymous reviewers of the original proposal, and Kees Vaes my editor at John Benjamins.

I thank the IUF (Institut universitaire de France) which has enabled me to dedicate myself completely to research for the last four years, and for the year to come.

This book can be used by students as well as academics and industrial researchers (linguists, logicians, mathematicians, computer scientists, software engineers, quality engineers, etc.) looking for new methodologies not only for natural language processing, but wherever language and quality meet, for safety critical applications whether involving professionals, the general public or both.

Prologue

"Man is not merely *homo loquens*; he is *homo grammaticus*" (Palmer 1975: 8). The concept of language is represented through diverse languages which need to be situated in the flow of time. We can look at these languages as systems composed themselves of systems and micro-systems from the point of view of the whole or from the point of view of the components knowing that all the parts are interrelated as the stars in the galaxy and the macrocosm.

As Pascal said:

> But the parts of the world are all so related and linked to one another, that I believe it impossible to know one without the other and without the whole...Since everything then is cause and effect, dependent and supporting, mediate and immediate, and all is held together by a natural though imperceptible chain, which binds together things most distant and most different, I hold it equally impossible to know the parts without knowing the whole, and to know the whole without knowing the parts in detail. (Pascal, Pensées, English translation 1958: 12)
>
> Les parties du monde ont toutes un tel rapport et un tel enchaînement l'une avec l'autre, que je crois impossible de connoître l'une sans l'autre, et sans le tout... Donc toute chose étant causée et causante, aidée et aidante, médiatement et immédiatement, et toutes s'entretenant par un lien naturel et insensible, qui lie les plus éloignées et les plus différentes, je tiens impossible de connoître les parties sans connoître le tout, non plus que de connoître le tout sans connoître en détail les parties. (Pascal 1670: vol 4. 111–112)

Theory, methodologies and applications will meet here. From macro to micro or from micro to macro, which to choose, and how to represent what we call systems and micro-systems, components and the whole in the context of language and languages? We will start with our galaxy or languages as macro-systems and see why we need a micro-systemic approach.

Table of contents

PART 2. Modelling the norms

Introduction

This book presents a way of considering language and languages in a similar way to observing other natural phenomena such as planets and the universe (or, rather more modestly, galaxies). Two approaches are possible, either one tries to analyse the phenomenon as a whole, or one tries to delimit some particular object, say a planet as a microcosm, in order to observe it in the context of the whole as a macrocosm, which in our example is the universe. In like manner to the telescope for observation in astronomy, one uses the microscope in biology. Whether it be the telescope or the microscope, we observe that both are based on the same science, that of optics.

As when we view the universe through a telescope, we think not only that language is too 'dim' and furthermore complex to be analysed in its entirety, but also that modelling the way the parts composing a language are interrelated is in reality very difficult.

Contrary to the current state of the art, in this book we present a way to look at language(s) in a microscopic manner which then leads to the macrocosm. As these microscopic parts are interrelated, we have to forget for the moment the traditional division into lexis, morphology, syntax and so on. Each of the elements of these micro-parts could in fact belong to different micro-parts, let us say rather micro–systems, whatever we want to demonstrate, to compose, to analyse or to generate.

We present an original way of decomposing a language and languages in order to bring into evidence norms, whether intra-language or inter-language. The notion of norm will serve as the basis for showing how a language(s) can be modelled by macro- and micro-structures to apprehend it(them) better. Our point of view, which combines linguistics and modelling by means of norms, results in a theory that is exploitable for very varied applications such as natural language processing and controlled languages which latter have the particularity of necessitating very high levels of reliability, and indeed for language applications in general where reliability is mandatory.

PART 1

System, language and its components

In this first part, System, language and its components, we firstly introduce the notion of systemicity independent of the discipline studied. We then see what linguists and grammarians have written about language as a system at different epochs. Keeping in mind that we will need to be able to describe our system of language and what it is composed of, and knowing that subsequently we will be talking of a system of systems, we are confronted with the difficulty of delimiting clearly the domains making up language. To this end we look at and review, adding as necessary our own observations, how grammarians and linguists have addressed this problem in terms of the way languages function. We address in particular language typology, lexicology, morphology and syntax, the various micro-components including the word, morphemes and syllables, parts of speech, semantics, and finally and of methodological importance, norm in language in respect of variously synchrony and diachrony, and good usage.

CHAPTER 1.1

The concept of system

1.1.1 System

Condillac (Condillac 1771: 1) wrote "*un système n'est autre chose que la disposition des différentes parties d'un art ou d'une science dans un ordre où elles se soutiennent toutes mutuellement*" – a system is nothing other than the disposition of the different parts of some art or science in an order where all these support each other mutually.

The notion of 'system' has resulted in much scientific investigation, and, depending on the disciplines involved and the authors' ideological leanings, the concept has received various interpretations. However it would seem that these interpretations share a certain common foundation to which it is interesting to draw attention. Over and beyond each school's own specificities, we draw up a list of the 'presumptive properties' of what it means to be a system as so expressed in a large number of scientific publications. This enables seeing, in the chapters that follow, in what manner a language and indeed languages are systems. These systems need to be brought to light by means of the images, which are often deformed, being provided by the 'telescope' (macro approach), but firstly after having been studied with the 'microscope' (micro approach).

It goes without saying that our presumptions are great in number and that such an initiative results only in an approximation which is evasive and which needs to be contently reviewed containing as it does multiple points of view many of which being disparate are even irreconcilable. In any case who can boast that they have examined all of these points of view?

1.1.2 Systemicity

For these reasons and without claiming to elicit some strict but Utopian common denominator, Yves Gentilhomme (1985: 35–36) sought to establish an inventory of presumptive properties of systemiticity which seem to manifest themselves in the scientific literature and not restricted just to linguistics, over and above those proper to each school or discipline. We provide in what follows a translation of this inventory:

There is a presumption of systemicity if and only if the object being studied possesses the following five macro-properties:

1.1.2.1 Identification: the object being studied forms a whole which is identifiable and can be isolated.

1.1.2.1.1 It ought to be able to be grasped by means of one and only one global idea;

1.1.2.1.2 we ought to be able to name it;

1.1.2.1.3 we ought to be able to distinguish it from another system;

1.1.2.1.4 we ought to be able to isolate it either materially or conceptually from its environment, whatever this be.

This last property means that there exists a real or imaginary frontier which is imposed or deliberately constructed, impermeable or permeable, well defined or vague, which enables deciding, at least in a sufficient number of cases, what reasonably belongs or not to the system. Moreover, having said that the frontier can be more or less permeable, this means that some interaction can exist between the exterior environment and the presumed system.

1.1.2.2 Structure, elementarisation: the object being studied possessing a coherent internal organisation, it is convenient to distinguish the different parts.

The supposed system which constitutes the object being studied can be decomposed in at least one way into more elementary components, that is to say, smaller, even more simple and sufficiently stable so as to be identified and inventoriable as sets whether defined or vague. These components maintain between themselves multiple connexions which can be coherently described notably in respect of relations and operations. Amongst these components, certain are declared to be primary, that is to say that in the framework of the analysis that has been undertaken we do not attempt to reduce them to more elementary components, whilst others appear as aggregates of primary components.

1.1.2.3 Interdependence, functionality: a subtle interrelation involving action and retroaction is enacted between the components and the links that they maintain, the existence (involving definition and determination) of the primary components being dependent on the secondary components as if these latter had for function establishing the existence of the former and vice-versa. Thus for some observer outside the system, it appears that this latter is finalised. The system constructs itself progressively as we observe it. The system can be called into question, reorganise itself completely, or cease to be considered as a system.

1.1.2.4 Originality: the whole does not reduce itself to an amorphous set of parts.

1.1.2.5 Persistence: the system ought to be able to be identified, conceived, distinguished, and named for a certain duration during which it possesses its own coherence and a particular organisation.