Methodological Foundations of Lexicon Building

Gabriel G. Bès

Methodological Foundations of Lexicon Building

Gabriel G. Bès
GRIL (Groupe de recherche dans les industries de la langue)
Université Blaise-Pascal, Clermont-Ferrand

Report for project LRE 61034 DELIS, revised version, July 1995.

Abstract

The present paper on lexicon modelling relates to the issue of building a generic and reusable lexicon. It is founded on the basal idea that there cannot be any adequate formalization or modelling without careful and explicit description of data. Here, data are lexical semantic data.

§ 1 sketches an overview of the domain and § 2 provides terminological and conceptual clarifications. It is noted that there is no explicit agreement on what semantic data are. A set of requirements the data must meet is defined, such as being accessible to actual and realistic observation, being intersubjective, being independent of any particular theory, etc.

§ 3 is the central chapter of the paper. It defines a conceptual specification of a constructive reusable generic lexicon. First, a series of statements on observational semantics are presented: on the paradigm (precisely that of empirical science), on the syntactic descriptions of natural language expressions (a strict and narrow definition), on the description of real actual worlds, and on the observer (the indispensable black box which turns factual observations into data descriptions). These statements are basic assumptions the reader should accept before going any further. Next, two propositions on semantic modelling are made, each with explicit assumptions on observational semantics (including the observer's knowledge and capabilities), specific data and specific requirements for theories on these.

In the first proposition, the observer uses a set of three predicates on sentences (syntactically well-formed, constructible, deviant), yielding a specific data set. As linguistic theories are not much explicit to what their target is with respect to this data set, strong difficulties are to be expected in the specification of semantic information in the lexicon, as well as a low level of intersubjectivity of data. From the observation of a few examples, it is argued it is impossible to model any lexicon semantics within the paradigm of empirical sciences in the terms of this (commonplace) proposition.

In the second proposition, the observer has in addition the ability to see that two expressions have or haven't an identity of interpretation. This allows to partition into two sub-classes the problematic class of sentences of the first proposition (i.e. well-formed, constructible, deviant sentences), depending on whether there is or not a well-formed, constructible, non deviant paraphrase. When all conceivable paraphrases are deviant, the problem is ontological, when a non deviant paraphrase exists, the problem is relevant to linguistics. Lexical semantics are thus distinguished by a principle from encyclopaedic semantics.

DELIS - LRE project 61-034

METHODOLOGICAL FOUNDATIONS OF LEXICON BUILDING

Gabriel G. Bès

GRIL
Université Blaise Pascal Clermont II

34, avenue Carnot F-63037 Clermont-Ferrand Cedex
Phone +33 73 91 54 44 / Fax +33 73 40 63 99

July 1995

SUMMARY

PREFACE

The present paper is the revised version of the GRIL's contribution on lexicon modelling. It is founded on the basal idea that there cannot be any adequate formalization or modelization without careful and explicit description of data, these in turn not being natural or well-defined by common sense. Data, in particularly data intended to be reused, must be methodologically specified : they are not natural things. If stored data are going to be subsequently offered, an explicitation must be given of the objects the data express and evaluation of their fiability must be specified. Empirical theories are tested in relation to data : collected data must thus be specified describing also the range of theories that in principle must account for them. This contribution has been thought with these general ideas in mind; these guiding principles are intended to be sine qua non requisites for anybody who wants to specify reusable data and in particular lexicon reusable data. The contribution sketches thus the foundations of a general methodology for building in a principled way reusable lexicon data basis intended to be something more than a collection of 'feelings' or intuitions more or less spread.

The contribution benefits largely from previous work of the GRIL on lexicon matters, noticeably work done in the GENELEX project and using the LADL tables, from previous writings of the author, some ones in collaboration with other authors and from a now rather extensive revision of the work in the field.

It is thought that the DELIS members know what is the general thinking of the author, expressed in previous contributions and in oral discussions in DELIS meetings, on the conceptual value of the frame adopted in the DELIS project. In this contribution, the target is not to revise or add anything to this frame. It would have been a senseless task. The targets are rather the definition of the overall and general requirements which should be met by any writing pretending to the status of *theory* on lexicon semantics, and the specification of concrete propositions in order to capture data which are intended to meet these requirements and that a range of theories must account for. The role of general requirements is not to impose any solution. It is rather to settle some rationale to discuss on semantic matters.

The contribution also benefits greatly from meetings where questions on lexicon semantics have been extensively treated : the Daghstuhl seminar of 1993, the Copenhague summer school of August 1994 and the RANK-XEROX meeting of October 1994. Work and courses by Kamp, Bierwish, Briscoe and others conviced the author that what he was saying for some time within the DELIS project with no success, was interesting enough to be said in the clearest and most explicit way he had the skill to produce.

Criticism of the propositions are welcome. In any case it is thought that we all agree on the fact that *two wrongs don't make a right*. It is not because the enterprise here sketched presents many mistakes and/or errors, that the present general picture on lexicon semantics will become attractive.

Last and not least, the contribution benefits from personal discussions with Paul Gochet, Jacques Jayez and A. Braasch during the Copenhague meeting, from an internal seminar of the CRYCIT linguistic team (Mendoza, Argentina, September 1994) where the author was invited to present his views and received challenging feed-back from V. Castel, D. Rossi and L. Paris, from internal discussions in the GRIL with J. Rodrigo, who is working on Spanish verbs, with M. Emorine who worked on the control of the lexicon of controlled languages in relation with AEROSPATIALE DAV (Toulouse), with K. Baschung, who has an inner view and practice of the descriptive apparatus of the DELIS project, with S. Aït-Mokhtar who is working on a powerful tool of text annotation which, among others, is intended to be used in the extraction of lexicon information and with C. Dafniet who helped, besides, in a better formulation of the written English. As always, the responsability is the author's alone.

G.G.Bès
July 15 1995

Chapter 1

PRELIMINARIES

General objective.

This contribution is on lexicon matters : subcategorization, lexicon semantics, definitions and the like. These questions are related to the idea of 'building', i.e., in some intuitive sense to the conditions that must be met to pass from one state of the lexicon to the other, and to the ideas of reusability and genericity. This contribution is thus on reusable, generic and constructive lexica but it is impossible at this point to state more precisely its content. The issue will be stated more accurately at the beginning of chapter 3. The objective of this Chapter 1 *Preliminaires* is to try to justify why this is so.

Difficulties to state the problems.

Lexica are part of the abstract system underlying natural languages. Lexicon issues relate to more general issues on the relation between symbolic systems -i.e. systems which are presented with some kind of notation- and something that is outside the symbolic system and that can be either another symbolic system or another object that can be called world or reality. This question has been studied from years by linguistis (structuralists, generativists, computational), dialectologists, artificial intelligence scholars, psychologists, logicians, philosophers of science -not to talk of other kinds of philophers-, semiotists and litterary critics. Lexicographers, people who produce actual lexica, besides producing them, had also express their ideas on this point.
It is thus unsurprising to meet with terminological difficulties. These are at least fastidious and at worst misleading. Working on the underlying concepts on which the terminology is built explicitly or not it is not impossible to clarify matters and is frequently revealing to do so. The profit may be to dissolve some spurious controversy (cf. Katz citation).
It is also unsurprising to meet with some definitive opposite statements on some crucial point. These two kinds of difficulty may be illustrated by the following two exemples, that illustrate them in the order.

Example 1. The word *model* . GB model and truth conditional semantics.

Example 2. Statistics and induction from corpora. Positive and negative evidence. Corpus based lexicography.

Historical perspective : streams, schools and chapels.

Mutual ignorance.
Example 1 .The bibliographies of Sowa, Mel'cuk and Jackendoff.

Example 2 . Model theoretic semantics and philosophical semantics.
Example 3 . Post-Chomsky models and lexicography.
Exeample 4 . CEE -including EUREKA- projects on lexica.

On the other hand, there is an enormous amount of references, each domain, stream or chapel is developing independently, probably assuming that the others do no exist (see Calzolari and the GRIL bibliography).

Some historical-conceptual organization of the main streams of the field.

Structural linguistics.
North America : Sapir, Bloomfield, Hockett, Pike, Harris, Gleason
Praguian : Jakobson, Prieto, others
French : (a) Pottier, Greimas ; (b) Gross, LADL tables
English : Firth, Lyons
Glossematics : Hjemslev, Siertsema
Generative linguistics : (a) Katz,Weinreich, Jackendoff, Wunderlich ; (b) Gruber, Fillmore, Lakoff, McCawley
Meaning-text model : Mel'cuk, Wierz....
Terminology and lexicography : Rey Debove, Sager
Philosophical semantics : Frege, Quine, Russell, Carnap, Putnam, Strawson
Truth conditional semantics : Montague, Thomason, Dowty, Chierchia
Dynamic semantics: Kamp, Heim, Stolckoff
Lexicon projects : Delis, Multilex, Eurotra 7, Genelex
Linguistic models : GB, GPSG, LFG, HPSG, Gategorial grammar
Artificial intelligence : Kayser
Psycholinguistics : Fodor, Garrett
Others : Pustejovsky

Notions.

Syntax, semantics, pragmatics in general theory of languages.

Meaning postulates, paraphrasis, analytic meaning, synthetic meaning. Definitions. Extensional and intensional meaning.

Compositional semantics. Truth values, truth conditions. Informational semantics.Truth conditional semantics. Model theoretic semantics. Entailment, implication.

Subcategorization, lexical meaning, syntax, semantics and pragmatics in linguistics. Thematic role, lemma, collocations. Predicate structure. Event structure. Time and aspect. Lexical entry, lexicon knowledge base, lexicon data base. Lexicon rule. Lexicographic definition; terminology.

The spirit of the contribution

The matters of the contribution relate to the issue of building a generic and reusable lexicon. This issue of lexicon building must be situated in relation to what is known today in the fields mentioned above. Several examples will clarify this point.
Example 1. In one way or another, the semantics of a lexicon entry has something to do with definitions. Even if the contrary has been decided i.e. that it has nothing to do with definitions, any assumption on the issue must be explicitly spelled. It is highly desirable to discuss and justify the the assumption. If the semantics of the lexicon has something to do with definitions, it has something to do also with the analytic-synthetic controversy. And it is impossible to act as if people as Quine, Carnap, Putnam, Kripke, Wittgenstein and Katz had never exist. What is the controverse ? What is the state of the art of the controverse?(Or what is supposed to be the state of the art ?). What are the consequences -if any- to the issue of lexicon building?

Example 2. In many writings the notion of 'thematic role' is used. The models used in truth conditional semantics manage to relate argument positions in predicative expressions to n-uples in the world. What is the mapping of a 'thematic role', if any?

Example 3. Dowty explores several accounts of the notion 'time' . What looks to be the same notion is used by Kamp, Jackendoff, Pustejovsky and others. It seems that depending on the extra-language notion of time adopted, the account of aspectual features of lexicon semantics changes. Does it mean that aspectual linguistics is ontologically dependent and, if so to what extend?

Example 4. In some writings there are occasional appeals to a distinction between (lexicon) raw facts or data on one hand and explanation of these facts or data on the other. Is it possible to transfer the distinction explanation/description into lexical matters? What are the conditions of adequacy of lexica, if any? In relation to this crucial point : are there different adequacy conditions for different kinds of lexica (reusable lexicon, domain dependent lexicon, lexicon in a particular grammar model) ?

As clarification is one of the objectives of the contribution, general formulations are often illustrated with detached examples, as in the above text. Detailed references to classic books will often be introduced, as well as revealing citations. Intuitive formulations of the ideas will be presented previously to formalized statements. Natural language explicit pre-formal presentations are often used : there are intended to be a guide to the interpretation of more formal statements. Clarity is thus sought even at the cost of some redundancy.

Brief description of each chapter.

The enterprise is thus on lexical matters, but it tries to relativize each matter in relation to different streams, traditions, fields or chapels. For this doing, previous terminological and conceptual clarifications are needded. These are undertaken in chap. 2. The clarification of chapter 2 allows to settle in the same chapter what the requirements on lexicon semantic data are thought to be. Chapter 2 ends coming back to streams, schools, chapels and previous writings and tries to establish the balance of existent and no-existent answers to questions previously raised . Chapter 3 - the central chapter of the book - restates formally the goals of the book and presents a conceptual specification of a generic lexicon. By the way, issues on implementation are raised, but carefully distinguished from conceptual and formal explicitation . Lexica must be related to corpora, both for corpora feeding and for corpora testing. Chapter 4 tackles this crucial issue, crucial because, among other things, the conceptual specification of the previous chapter is intended to define specifications which must be used and tested. In chapter 5, the generic lexicon is related via its semantics to domain lexicon, to terminology and to encyclopaedia. Chapter 6 concerns perspective propositions, which relate the propositions presented in the contribution with other work in

the field, and evaluative propositions, which develop the issue of reflexive knowledge od lexicon data bases.

Chapter 2

METHODOLOGICAL BACKGROUND

The epistemological pattern of an empirical science.

Brief recall of the neo-empiricist pattern. Hempel, Bochenski, Popper, Bunge. Theory, description, data. Falsification.

Linguistics as an empirical science.

A language study paradigm. From structural linguistics to Chomsky 1955. From Chomsky 1955 to the present days. Internal and external adequacy. Levels of external adequacy.

From claims to actual work. The continuous appeals to parallels with physics in linguistic work : Chomsky, Sag, Pollard, Jackendoff. Effective theories distinguished from jargonized texts self-claiming to be theories.

Negative examples extracted from current research in lexicon semantics :

- Pustejovsky's work. The so-called 'qualia structure' theory.
- The Genelex project : from the goal of a standardized model to a toolbox of formalisms.
- The admission by several intervenants at the Rank-Xerox meeting (Grenoble 1994) of the inexistence of clear semantic data.

Suppose that T = specification of reusable lexicon data base in their semantic aspects.
One of the basic assumptions underlying the present contribution is :

- we all have some idea of what T means
- any pair of us don't have the same idea

and this because the state of the art can be summed up succinctly in the following way :

- there is no explicit agreement of what semantic linguistic data are

Fixing the frame of the present contribution

• **Methodological assumptions.** Reality, real data. Explicitation. Formalization (levels). Symbolic data and real data. The modelization of the Observer (= Obs).

• **Terminology and notation.** Pre-theoretical notions (pattern, object, entities, etc). Assumed definitions. Formal entities (lists, sets, variables), linguistic entities (subcategorization, syntax and semantics mapping), logical entities (inference, truth values, entailment, material implication, implementation entities (factorization of the information, rules, primitive information, symbolic data base, symbolic data structure).

The need of semantic data .

Besides the requirements on data coming from methodological general assumptions on empirical sciences, there are at least four specific motivations requiring the systematic capturing of semantic data.

(i) Motivations from Eurotra 7

• Abstraction of abstraction = data

Reusability is not defined on grammars but on KB about *data* . Information can be extracted from KBs in order to build grammars (and other linguistic knowledge sources) accessed in actual processing systems. The consequence is that you cannot talk of reusability if you don't have an explicit definition of data.

(ii) Motivations from EAGLES

• The evaluation of the functionalities of NL processing systems and their underlying formalisms requires explicit description of data coverage.

• Important improvements have been done in the notation of syntactic dependency data required for compositional semantics. Systematically noted texts are used in systematic testing of NL processing systems (cf. MUC).

Economic pressure on technology evaluation is actually dissolving the mythical, damageable and erroneous idea according to which any linguistic data is dependent to a particular theory.

(iii) Motivations from the primitive definition of the DELIS project.

In the original planning of the DELIS project, an explicit 'cookbook'was required in order to feed the lexicon data base. Furthermore, in technical presentations of the project, the 'implementation cycle' is explicitly and repeatedly mentioned. This cycle is parallel to the hypothetic-deductive pattern of empirical sciences. In both cases falsification by means of data is required in order to reformulate general hypothesis or existing implementations : without falsification no improvement is possible. And, without data, no falsification is possible. More concretely : you cannot be consistent and affirm both that you are using the testing cycle **and** that you don't have any method to assess dictionaries (or any lexicon data base).

(iv) Motivations from the intended use of a source of reusable knowledge base

A reusable data base is intended to be used. This use minimally requires explicit description of the data contained in the source. In the lexicon domain, this minimally requires from a lexicon data source to be specified by an explicit definition of its content, i.e. to have explicit reflexive knowledge on its own data.

Special requirements on semantic data

Data must be observable and described in uniform way by qualified observers, i.e. data must be intersubjective. Besides this general requirement on data coming from the general pattern of empirical sciences, semantic data meet with two supplementary difficulties.

• All empirical sciences (physics, biology, chemistry, etc) express knowledge about RAWs (= Real Actual Worlds) as much as common sense does. Linguistic semantics relates to the same RAWs than the other knowledge sources. Has the knowledge encoded in linguistic semantics to be in principle distinguished from encyclopaedic semantics, the knowledge coming from the other sources? If so, on which bases?

• At any arbitrary state, a linguistic reusable lexicon will never be complete or definitive. Data must be specified so that their integration into the successive states of the constructive lexicon will be possible.

Despite difficulties, it is not unreasonable to think that there must be some way of talking about semantic lexical data in a uniform and explicit way and to distinguish them from theories about lexicon semantic data bases. Exemplary work by Dowty points in this direction.

The challenge

Assumption : data of lexicon semantics concern the relation between observable NL expressions and RAWs.

Data must be :

• accessible to actual and realistic observation
• intersubjective
• independent of any particular grammar model or theory
• integrated in a particular scientific paradigm
• reusable by any particular grammatical model in the explicited scientific paradigm
• expressible in some representation language
• integrated into the states of the constructive lexicon with explicit definitions of the transitions between the states
• related and/or distinguished from encyclopaedic knowledge

Last but not least :

• data must be significant, i.e. they must concern observations which are ordinary contended to be interesting in the writings of the field.

The state of the art

The state of the art can be synthesyzed by the following five kinds of nonexistence.

• Nonexistence of consensual agreement on semantic data

Conceptual assumption : Morris distinction on syntax, semantics and pragmatics; *semantics* being the relation between language perceptible expressions and the world. All the following positions are observed in the field :

•The relation is not justified at all.
•The relation is justified in terms of :
(a) Platonic objects (Katz)
(b) Output of effective parsing (Fodor, Garrett,)
(c) Cognitive content of the mind (Fodor, Langacker, Jackendoff, Lakoff)
(d) Restricted situations of the world (stimulus meaning, Quine)
(e) The relation is an assignement function F and the world is a set of symbolic objects organized in sets and n-uples (model theoretic semantics)

Furthermore we have internal subdivisions crossing the others. The relation between language expressions and the world
• is mediated by some intermediate structure into which language expressions are translated and which
• is in principle necessary (logical form in GB, Katz, Kamp's DRS, IL in GPSG, Mel'cuk, Sowa's conceptual graphs)
• is there as a convenience but is not necessary in principle (Montague's UG and PTQ)
• is not mediated by any intermediate structure (dynamic semantics) which is explicitely denied (Quine) or considered with suspicion (Harris, Gross).

The use of a wide variety of expressions related to meaning observations (*out, it is not good, it is logically absurd, analytic statement, semantic contrast, deviant, felicitous, grammatical, acceptable, semantically deviant,* and surely some others) is another good indication of the nonexistence of a consensual agreement on semantic data.

• Nonexistence of candidate data fulfilling the requirements of accessibility on semantic data.

• semantic data are not platonic objects.

• outputs of effective parsing procedures and cognitive contents of the mind for deontological reasons are not in principle observable (cf. Kamp) and for practical reasons nothing says that they will become so some day.

• with the symbolic objects used in model theoretic semantics (in any of its patterns) the main problem to solve is assumed to be solved Furthermore, the possibility of calculating extensions from intensions lies on the asumption of fully specified extensions for each predicate in all worlds.

"Another kind of information we must supply in interpreting L is the intended *intension* (or meaning) of each predicate of L. To do this for a predicate P we should determine, for each point of reference i, the *extension* (or denotation) of P with respect to i. For example, if the points of reference are moments of time and P is the one-place predicate 'is green', we should specify for each moment i the set of objects to be regarded as green at i; if P is instead the two-place predicate 'is married to', we should specify for each moment i the set of ordered pairs $< x, y >$ such that x and y are to be regarded as married at i ."

(Montague, 'Pragmatics', in *Formal philosophy* (1974), p. 98)

• Nonexistence of a principled distinction between encyclopaedic semantics and linguistic semantics.

Again, a variety of positions can be recognized :

• the distinction is denied

• the distinction is
  • explicitely assumed but not defined (Mel'cuk)
  • is defined in terms of platonic objects (Katz)
  • is intended to be defined but in theory dependent terms (Bierwish)
  • is implicitely assumed but not defined (frame semantics, Fillmore)
  • is not required in model theoretic semantics.

• Nonexistence of reflexive knowledge of lexicon data bases

No actual lexica data source and no model for any lexica data source worked out in any project (including LADL Tables, Genelex, Acquilex, Delis and all the others) is presently described with its reflexive knowledge

Chapter 3

CONCEPTUAL SPECIFICATION OF A CONSTRUCTIVE REUSABLE GENERIC LEXICON

Chapter 3 is the core chapter of the contribution. Its underlying guiding ideas are :

- no semantic lexicon model outside the hypothetic-deductive pattern of empirical sciences; i.e; no model without carefully specified data and without rigorous formalisms expressing them.
- description of data must be carefully distinguished from theories about data
- no formalism without a methodology for observing data which must be expressed by the formalism.
- specification formalisms must be carefully distinguished from implementation formalisms.

3.1 Restatement of the goals of the contribution.

Technical restatement of the goals of the contribution in terms of the definitions, distinctions and requirements introduced in the previous chapter.

3.2 SOS : Statements on Observational Semantics

Specification of a finite set of Statements on Observational Semantics (SOS) from which it will be possible to define three kinds of Propositions defining the data aspects of the overall model.

SOS express the assumptions which must be accepted in order to specify the propositions that will specify the relation between semantic data and theories of natural language. In other terms : if you accept this and those statements, it makes sense to define this and that requirement on the semantics of the entries of a lexicon. For example, if you don't accept that linguistics are an empirical science, dont bother about what it is said in a proposition which has as one of its essential goal the one of testing. But if you refuse this kind of validation, please spell what is your own one. If you don't spell any one, you must know that, in any

other science than linguistics, people having this kind of behaviour will be strongly ill-considered.

3.2.1 General view on the statements

The statements are presented in the following pattern :

i Assumptions (Derived or axiomatic)
ii Definitions
iii Corollaries

Intellectual inspiration : Bloch A set of postulates for phonemic analysis.

The statements are on :

• the general features of the paradigm, including the scientific methodology adopted, the relation between NL expressions and RAW's and the relation *account* between theories of NL expressions and NL expressions
• the Observer and his outputs
• the description of RAWs
• the syntactic (in Morris terms) description of NL expressions

3.2.2 A summary of required statements :

On the paradigm

• Morris definitions of syntax, semantics and pragmatics.
• Scientific methodology of empirical sciences.
• Sentences have truth values.
• Sentences truth values are evaluated compositionaly from the truth conditions of their *well formed syntactic sub-strings*. The composition is specified in terms of *syntactic dependencies*.
• An explicit function F (relation?) assigns recursively well-formed syntactic strings to aspects of RAWs from the bottom up, beginning with the strings which are GP (see immediately bellow the explicitation of 'GP') .
•The assignment of a GP to some aspect of RAWs is exclusively conditioned by the requirements expressed by SeI in < GP, SyI, SeI, -LC >on the aspects of RAWs to which GP is assigned (see immediately bellow the explicitation of 'LC') .

The consequence is that if Sei is empty any assignment is possible i.e. the GP can be assigned freely to any aspect of any RAW.
•An Obs is needed to evaluate the assignment relation.
•Data which theories on NL expressions account for represent NL expressions in some notation : data for theories are thus represented data expressions. A represented data expression, or to shorten, a *data representation* includes minimally :
(i) a signal representation solving the type-token relation between the perceptible aspects of different token expressions (this is ordinary named 'phonetic or phonic representation' and/or is presented as a string of printed characters).
(ii) a classification of expressions in some basic categories, as e.g; +/- grammatical, +/- ambiguous, etc. Let call these categories *judgements on NL expressions* . This classification is ordinarily spelled by specific symbols associated to signal representation : the '*' symbol is the notation mark widely used for the non grammatical judgement.

It is Obs who assigns representations to data (see below). More sophisticated are the judgements the Obs must perform, more skills must be allocated to him in order to accomplish them.
•A theory on NL expressions specifies *theory representations* and defines an explicit mapping between the theory representations it specifies and data representations. It is this mapping between theory representations and data representations which express the notion 'a theory of NL accounts for NL data'. The mapping must express which symbol(s) or structures must be assigned to which symbol(s)or structure(s) in the data representations. Ideally, there must be a bijective relation between the set of theory representations and the set of data representations such that for each theory representation with a symbol X there must be a data representation with the symbol Y corresponding to one particular judgement.

Comments

Previous statements on the paradigm introduce no revolutionary innovation with respect to current scientific or linguistic practice. Rather, the statements intend to explicit this practice. In any empirical science, data are needed, an Obs is needed, even if he is helped by sophisticated apparatus that enhance accuracy of observations when these are measurements. For long time, physics, biology and chemistry requested and still request an Obs with specific skills : he must be able to translate evenemential observations on RAW's into data representations. In each science, the evaluation of the average margin of error for different types of observations is an important issue. Errors, fluctuations and noise exist in data but they do not invalidate the whole pattern, inasmuch they are explicitely recognized and evaluated. We assume that even for the time being, situations in which sciences can dispose of sophisticated collectors of data translating physical stimulus into data representation without human interaction are more the exception than the rule. We conclude on this point saying that linguistics does not present a principled obstacle for not being considered as one of the so called hard empirical sciences.

On the syntactic descriptions of NL expressions

• NL expressions are strings of graphic primitives (=GP)
• A GP is a string of graphic symbols from a finite alphabet
• The lexicon is a finite set of quadruplets of the form :
<GP, SyI, SeI, LC >
where SyI = syntactic information, SeI = semantic information, LC =logical constant
Gp and SyI are not empty; SeI may be empty; the LC label is + or -.
• There are well-formed syntactic conditions on NL strings
(Here and elsewhere *syntactic* is used in Morris terms (see below); syntactic conditions can be exclusively formulated in terms of extensionally defined sets of quadruplets, linearity conditions on GPs, graphic symbols in GPs, existence requirements on GPs or sets of GPs in the strings and generally in terms of any kind of property relying only on internal properties of strings of GPs).
• Sentences are a subset of well-formed strings.
• Sentences are strings of well-formed (sub)strings.
• There are syntactic dependencies between well formed strings.
• A GP is a well-formed syntactic string.
• A syntactic space is a set of strings meeting syntactic requirements.

Comments

The assumption is that, here again, there is nothing (or very little) new with respect to current practice, with the exception of the strict use of 'syntax'.

The 'LC' symbol is there for eliminating entries as *and, each,or, no* which can be marked +LC and be semantically interpreted in their own way.

What is intended by 'syntactic' can be best explained with an example.

Reading the syntactic presentation of an artificial language, eg. some logic or some programming language, i.e. the clauses which express conditions on well formedness, it is possible to observe a common pattern of presentation. Firstly, different kinds of alphabets of graphic symbols are presented. These graphic symbols are presented ostensively or refering to an intensional class of typographic symbols; e.g; 'variables are represented by capital letters', syntactic symbols in the expressions are '(,)'; operators are : -->, &, ... " and so on. Secondly, there are basic rules of linear combination saying which kind of typographic symbol can combine linearly with which others; e.g; "if P is a variable and Q is a variable, P & Q" is a well formed expression. Finally, there are recursive rules of linear combination : "if E1 and E2 are well formed expressions, E1 & E2 is a well formed expression".

With this kind of definition, the solving of the type-token relation between expressions is entirely left to the skills of the user.

In this contribution, syntactic must be interpreted in the same previous strict sense : i.e. syntactic is any kind of representation which can be expressed either in terms of symbols in the signal representation, or in terms of classes of these symbols defined extensionnally (i.e. it is possible to fabricate any ad hoc syntactic class of lexicon entries by a simple enumeration of the members of the class, each member being identified by its GP) or intensionnally in terms of linearity relations between GPs or classes of GPs. A syntactic dependency and a syntactic rule are thus a relation that can be expressed between any two or more strings of syntactic entities as defined above and, in the case of rules, that has syntactic consequences. The subject function, inasmuch as it can be expressed in terms of obligatoriness, linearity relations and agreement, this in turn being defined in terms of cooccurence relations, is a syntactic dependency.

With this definition of 'syntax', we are not committed in the following to discuss if what is intended here as being 'syntax' matches or not with what is dummed 'syntax' in other writings or models of grammar. The general assumption is that 'syntax' and its derivates are rather very loosely used in linguistic writings. We are neither committed to discuss if a grammar must be 'syntacticly or semantically founded', whatever this may mean -if anything.

The use of syntactic with the previous spelled meaning will allow us to express limits on the phenomena that syntactic representations specified by any grammar can express. That is the basic idea is to state : if this kind of data is interesting, representations built in syntactic terms are either not enough or involve this and that difficulty or have these equivalent properties with representations which can be obtained introducing semantic considerations (to be defined).

On the description of RAWs

- Aspects of RAWs assigned by F to GPs of the form <... -LC> can only be intensionally described with descriptions of the form
  { X | such that Y }
- It is not possible to recover extensions from intensional descriptions
- n-ary predicates used in intensional descriptions are vague.

Comments

Again nothing new is expressed in the previous statements on RAWs. It is supposed because they are accepted as true, that anybody trying to evacuate metaphysical (i.e; ontological) presuppositions develop non intensional semantics in logical minded oriented writings (e.g. Quine) and to evacuate semantics from linguistics formal descriptions (Harris, the first Chomsky down more or less to 1965). In the other hand, people claiming themselves as mentalists (e.g. Jackendoff, Ross) never try to undestand why ontological presuppositions introduce perverse results. The present situation is thus one where people working with concepts and mentalism act as if truth conditional semantics never existed and do not bother about the crucial objection saying that inner states of the mind, where concepts lie, cannot be observed and that a science without in principle observable data is not a science. On the other way round, people introducing intensionality technically (Montague, Sowa) suppose a description of RAWs (see above Montague citation in chapter 2) which is as unrealistic as the one of mentalist people. What we are here trying to do is to catch insights coming from the two orientations, the modelisation of Obs being what it seems to be the indispensable cost to pay for doing so. An extraordinary lucid state of linguistics semantics was given by Manfred Kripka in his talk at the Copenhague Summer School of 1994.

On the Obs(erver)

The Obs represents a set of qualified human beings capable of :

(a) solving the type-token relation between perceptual aspects of NL expressions.

(b) learning and applying judgements on NL expressions in order to obtain data representations (i.e. judgements of the grammaticality type, sentence character of strings, etc) and being able to apply discriminatory tests to NL expressions.

(c) evaluating the account mapping between theory representations and data representations

(d) expressing the results of observations in some explicit language (i.e. producing effective data representations)

(e) evaluating the F function between NL expressions and RAWs via an intensional description of RAWs in terms of selected predicates

(f) eliciting from the Users and/or inducing from the Users's behaviour in speech acts their underlying judgements on consistency, on truth values and on truth conditions of NL expressions.

Comments

The basic methodological idea presented here is thus the following : Obs is the indispensable black-box required to turn evenemential phenomenic observations into data descriptions and to turn RAWs into represented RAWs. We don't know how Obs works but we assume that Obs can be trained to work in some controlable and consistent way, within controlable margin of variation. It is assumed that the introduction of Obs is the cost that must be paid in order to make operative model theoretic semantics , intensional semantics, dynamic semantics and in general any kind of semantics relating NL expressions to something outside them, which is directly or indirectly observable.

The allocation of skills to Obs will be done bellow gradually. In the gradual and constructive pattern adopted here increase of Obs skills will improve gradually; with each possible

improvement new kinds of data will become accessible but it may be that at certain points consistency of results will diminish.

Contrary to some to common practice, Obs is not asked to express his intuitions about the beautifulness, interest, goodness or the like of theoretic representations (this kind of judgements are left to the authors of the representations themselves) and does not decide for any theory or model of grammars if some particular data must be accounted in its 'syntactic' or 'semantic' component. He does not discuss neither about what are 'natural' kind of 'synactic' data for any theory of language.

3.2.3 Propositions

SOS furnish the assumptions which must be accepted in order to specify propositions. Propositions define the conditions of success of semantic modeling. SOS fix the minimal conditions required to talk seriously on semantic matters. Propositions try to talk seriously on semantic matters.

There are three kinds of propositions :

I Data-theory propositions
II Perspective propositions
III Evaluative propositions

I Data-theory (D-T) propositions relate :

- accepted paradigm
- type of data
- Obs knowledge & capabilities
- operational evaluation of Obs outputs
- syntactic spaces defined on NL expressions
- specification of SeI (Semantic Information in lexical entries)

The pattern of a D-T proposition is :

Given :

(i)
- paradigm X
- data Y
- syntactic spcaces W

(ii) any G (theory of language) accounting for data Y in the paradigm X within W

then :

• G accounting for data Y in syntactic spaces defined W must incorporate SeI of type Z in the lexicon.

• the accuracy of description of data Y will never be superior to score n

• the knowledge & capabilities of an Obs capable of specifyng data Y that, among others, are indispensable for testing any theory on them, must be Y'.

II Perspective propositions relate the present work in the field of linguistic semantics (including terminology) to SOS.

III Evaluative propositions state

• the coverage and fiability of lexicon KBs (reflexive knowledge)

• the fiability of any calculus incorporating SeI of lexicon entries

3.2.4 Data-theory propositions

D-T propositions model lexicon semantics. The underlying guiding idea is that there is no such a thing as THE semantics of a lexicon entry. Instead we have a gradient of possible semantics beginning with a zero level, the level in which there is no semantics at all and, by consequence, no requirements coming from GP on the assignement function F. Further levels increase SeI sophistication but at each incremented level there is a cost to be paid in terms of scores of accuracy of description. Possibly in the highest level you can account for the delightness of enjoying your favorite poet, but intersubjectivity scores presumely will collapse to zero.

A D-T proposition relativize the type of SeI associated to GP in terms of type of data considered in some well-defined syntactic space, given a paradigm stable (the one of empirical sciences).

We will consider in the following two D-T propositions (N°1 and N°2). Each one will be presented with the crucial SOS assumptions needed to apply it. For both, the considered syntactic space will be the sentence. Data and requirements of the theory on them will change in each proposition.

D-T proposition N°1 is intended to capture some features of the more or (rather) less explicit standard way of talking on lexicon semantics. We think it has many inconveniences.

D-T proposition N°2 proposes a reformulation of the kind of data and the skills of Obs and defines a general criteria allowing to introduce an incremental way of specifying SeI in lexicon entries.

D-T proposition N° 1

Assumed SOS on Obs :
Obs can apply the following predicates to NL expressions. The predicates are the primitives of the observational system :

+,- **Deviant**
+,- **Constructible**
+,- **WFSyE (Well Formed Syntactic Expression)**

Obs can elicit from the **User** the following intuitive judgements on expressions :

• **deviant expressions** : expressions presenting any kind of anomaly or strangeness
• **constructible expressions** : expressions being understandable as a whole from the combination of its parts; he/she is able to talk about the meaning of the expressions, can discuss about it, explain its denotation.

The assignment of the values to the predicate WFSyE is not achieved by Obs eliciting them directly from the User of NL expressions. The issue is much more involved. Its difficulties stem from the fact that -contrary to artificial languages - there are no well-formedness conditions a priori neatly defined for NL expressions. [Definitions are going to be presented here in order to specify different kinds of situations. The underlying idea is to try to extract from [-dev, +cons] expressions their syntactic conditions on well formedness and to project them into [+dev] expressions; *justified syntactic conditions* are the ones which are required by [-dev, +cons] expressions].

Applying the above predicates makes possible to obtain the following partition.

Partition 1 on NL expressions :

	+wfsye		-wfsye	
	+cons			
-dev	+dev	-cons	-cons	+cons
i	ii	iii	iv	v

Each class in the partition is illustrated by the following examples. A dubious example, i.e. an example which can be ranged in more than one class, is noted with *(?n)*, where *n* indicates the other possible class where it can be ranged.

i La fille aime le fromage.
La table aime la cire (?ii).

ii La table aime le fromage.
Colorless green ideas sleep furiously (?iii)
Jacques conduit le colis à la gare.
Un essaim de chiens courait dans la campagne.

iii Colorless green ideas sleep furiously.(?ii)

iv Un un un deux à.

v Un fille regardent le fromage.

The partition N°1 expresses data which Grammars must account for. This is expressed in D-T N°1, with its consequences.

D-T N°1

Given :

i • paradigm : empirical sciences
• data : partition N°1
• sentence space

ii theory representations specified by G maps biejectively with data representations which present one of the following partitions :
(a) partition N°1 above presented
(b) classes i, ii and iii of partition N°1 plus the union of iv and v
(c) clases i and ii of partition N°1 plus the union of iii,iv and v

Then :

• strong difficulties in the specification on SeI in the lexicon

• low level of intersubjectivity of data

Linguistic theories are quite generally rather not to much explicit and it is difficult to know what their tagets are (i.e. which one of the three possibilities a, b or c they are seeking for.. The crucial point here is the distinction between classes i and ii of partition N° 1, which appears in anyone of the three possibilities.

The underlying idea of this pattern is that each item must be associated with the complete properties attributed to its denotation by something which looks as 'common sense knowledge'. These properties are expressed by unary predicates (eg. +/- human, +/- artifact, +/- transportable, etc), sometimes called 'selectional features or restrictions'.

The two principal objections on this pattern are the absence of intersubjectivity of data and the insufficient expressive power of unary predicates.

• Non intersurbjectivity of data. Consider the example :

Un aveugle a conduit le camion

The previous sentence looks at first glance deviant and thus belonging to class II. Selectional features in the *conduire* and *aveugle* entries can settle the problem. The point is that, in October 94, newspapers reported seriously the denotation expressed by the sentence : it was an anomaly of the Italian system of paying pensions to handicapped people.

• Intensional unary predicates do not suffice for the expression of properties.

In many writings on lexicon semantics, it appears that some object is such and such. In Jackendoff writings, for example, some object (say *road*) is a path. Here the general point is that an object may be something in relation to some others but not in relation to all the others.

A *road* is a path in relation to humans, but a twig may be a *path* for an ant and not for humans.
In the same vein, in Pustejovsky's terms, some object has such or such telic role (directly imported from Aristotle). This is part of the so called 'theory' of Qualia structure :

Qualia structure:
Telic role : unary predicate

Ex. : book
Telic role : reading

As soon as you leave the previous and all the time repeated example, it becomes apparent that on one hand you do not know what are the telic roles of many things (In *I enjoy my grandchildren,* what are the telic roles of *my children* ?), and, on the other hand, that telic roles of things, if any, are relativized to other things and circumstances or events where these things appear. Consider the following examples. If *telic role* has some descriptive content at all, it is clear that the water is used for achieving three different goals in the three following examples :

Peter enjoys the fresh water in the glass
Peter enjoys the fresh water in the swiming pool
The trout enjoys the fresh water

The general problem stems here on the recursive, bottom up evaluation of F (if this is accepted in SOS; if not, you must discuss the question from Frege up). Observe :

(i) On a deplacé la montagne.
(ii) On a deplacé la plume.

Expanding the example, each one can either stay non deviant or become deviant :

(ia) On a deplacé la montagne en utilisant plus d'un millier de pelles mécaniques et des centaines de camions pendant des dizaines d'années.
(ib) On a deplacé la montagne en utilisant la petite pelle de mon petit-fils.

(iia) On a deplacé la plume en utilisant la petite pelle de mon petit-fils.
(iib) On a deplacé la plume en utilisant plus d'un millier de pelles mécaniques et des centaines de camions pendant des dizaines d'années.

The conclusion on this point of the present contribution is clear : it is impossible to model any lexicon semantics within the paradigm of empirical sciences in the terms of D-T N°1.

D-T proposition N° 2

D-T proposition N°2 is in some sense less ambitious than D-T proposition N°1 and more 'language founded'. We claim that D-T proposition N°1 attributes to natural language jobs which do not pertain to expressive tools : you do not encode in an artificial language the way of deciding if what you are saying is intelligent, true, efficient or whatever in relation with its denotation. Instead, D-T proposition N°2 does not have this goal.

The following ideas on NL are commonly accepted :
- arbitrariness : the specification of an assignment function F is needed to relate natural language expressions to their denotations
- effability : any possible or conceivable denotation in any world can be expressed in some WFSyE expression of any natural language.

These ideas belong to the underlying ones guiding D-T proposition N°2. The second one, in an intuitive sense, says that you can say absurds things in natural languages without any problem : mountains can be displaced, people can be at two different points of space at the same time, tables can think. It is in this sense that neither selectional restrictions nor definitions are an internal problem of language, even if it is a problem related to language that it must be accounted for in dictionnaries and other kinds of lexica.

The first one captures in a model-theoretic perspective the 'arbitrariness' of Saussure's sign (or Plato's Cratyl). But this formulation of *arbitrariness* will be changed into the one of *restricted arbitrariness* , here expressed in a non technical way by the following :

• from the exclusive syntactic (in Morris terms) description of linguistic forms in natural language expressions, it is impossible to specify the assignement function F.

Restricted arbitrariness introduces an important restriction to the commonly accepted idea of *arbitrariness* , even if it preserves the essential of Saussure's intuition,.recognizing the need of the specification of an assignment function F. Intuitively, it says that any natural expression (string of Wfsye) cannot be assigned to any arbitrary denotation by some F. It is thus possible to talk of *restricted arbitrariness* : when compositional semantics intervenes, you can always say absurd things but you cannot say them anyway. There are restrictions coming from the description of the denotata which allows you to have or not the interpretation itself in ANY WORLD.

This point is intended to be one of the cue points of the methodology here presented, with a departure from the Montague and model theoretic treatment of the same question.

The basic idea of F is to interpret language expressions mapping them into entities in some well defined symbolic world (data structure), language expressions being either syntactic theoretical representations (minimally, strings of GPs) or expressions of some logic language translated from them (with heavy restrictions on the biunivocity of translations in Montague). Sometimes general restrictions are introduced on the interpretative mappings of language expressions into the world entities, coming directly or indirectly from syntactic categories of the NL expressions (e.g. some syntactic category of NL expressions -e.g.verbs, nouns or other morpho-syntactic category- must be mapped into specific denotational categories of objects in the world; on the projection from NL expressions into the world, see below).

The departure from this position consists in the following : some syntactic objects can be mapped to specific objects of the represented RAW inasmuch other syntactic objects standing towards the first ones in some specific syntactic relation are mapped to some specific objects or categories of objects in the same RAW. That is, the core of the problem does not consist in the description of RAWs only (or, in Montague' terms, possible worlds) but (i) in the introduction of requirements in F itself and (ii) in formulating these restrictions in terms of symbols in the domain of F which are not syntactic, these symbols requiring a F mapping into specific denotations.

Quick examples on the general question of *pre-emption* and related matters can clarify the issue (for more details cf. Bès & Lecomte). Observe :

(i) Jacques a tué Pierre
(ii) Jacques a tué la valise
(iii) Jacques a descendu la valise
(iv) Jacques a descendu Pierre

Depending on subjective and individual opinions on animist matters, (ii) will be judged 'anomalous', 'deviant' or 'false in all worlds' or whatever label you want, or not. But in all cases, for doing so, you have to interpret (ii) as expressing the same relation between*Jacques* and *la valise* than the one expressed in (i) between *Jacques* and *Pierre*. The sentence (iv) can

be understood as in (i) or either as saying that *Jacques* changed the position of the case into the low direction. The sentence (iii) cannot be interpreted as in (ii) : the only possible interpretation is with a change of position. In French, you can kill a case with a *tuer* form, you can not kill it with a *descendre* form. In a some more general formulation : certain forms (to be precisely defined in terms of syntactic internal properties) admit any arbitrary interpretation, even absurds ones given some ontology, and others do not. The first ones can express absurdity, not the latter. In the subcategorization frame of *descendre,* when certain denotations are used in the instantiation of the object, some particular denotation of *descendre* is pre-empted.

There are no problems with French/English translation in the following pairs :

(i) Pierre gagna la couse
Peter won the race
ii) Pierre gagna beaucoup d'argent
Peter won a lot of money

If we add

Pierre gagna Paris

it is impossible to translate it as

Peter won Paris

This expression is not understood at all (i.e. it is not constructible in our terminology) or it is understood with some kind of meaning related in a not specified manner to the ones in the firts two pairs. The corresponding French expression is interpreted as being in some sense equivalent to *Pierre alla à Paris,* while the interpretation of the corresponding English expression is excluded for *Peter won Paris*. We are thus in the same kind of pattern that the previous example : depending on a particular form of a particular language, and on a particular denotation of a syntactic complement of this form, it is possible to map by F the form into a particular denotation or not. This looks as having nothing to do with ontologies about RAWs. They are restrictions on the mapping function F of French and English in any world. In the two languages (effability principle) it is possible to talk about peoples going to towns but it is a peculiar fact of French that it is possible to express this by using the same GP than the one it is used to express that there are won races or money .

Analogous pattern appears in diathesis. In the following two examples, the denotation of the dative complement express the denotation to which belongs the denotation of the object complement :

Pierre a volé un diamand à Marie
Pierre a volé un diamand au bijou

If we consider now the following two examples

Pierre a volé Marie
Pierre a volé le bijou

we have as data on them that the first one, in some romantic situations at least, can be understood as Marie being kidnapped, or either, more commonly, that something has been stolen from her. This second interpretation is completely excluded in the last example. There are in French some manifestations of reduced dative shift, but they are summitted to denotational constraints . And again, this appears as having nothing to do with the state of

affairs in the world. You can be perfectly well understood in French if you want to steal something to a jewel, but you must say so in the dativ way :

Pierre a volé un diamand au bijou.

As soon as you recognize that denotata descriptions introduce themselves as variables of the assignment function F, descriptions of denotata are not needed only in the description of the codomain of the function. There are some descriptions which are needed in the specification of the domain of F itself.

This point illustrates another cue of the propositions of this contribution on the modelling of lexicon semantics There are data concerning the relation between natural language expressions and the world that are indifferent to particular ontologies and that, under an explicit modelization of Obs can be collected by Obs in a consistent way. But the accounting of data requires the use of a descriptive apparatus founded in some ontology. And this is not a pre-fabricated one by some semantic guru; it becomes an hypothesis that, as any effective hypothesis in an empirical science, must be justified; furthermore it leaves the place to define formally the equivalence properties or not of competing hypothesis.

The kind of chosen and specified data selects the descriptive predicates of denotata that must be used in the specification of SeI in lexicon entries. It leaves open all the other you may use freely inasmuch the distinctions introduced by the ones used in the description of SeIs are preserved. It thus emerges a double sided picture of denotata descriptions : on one hand, the ones motivated by the accounting of some kind of well defined data; they are specified on SeI in entries. On the other hand, any expansions of the specifications in SeI; these can present individual variations inasmuch as they do not modify the specifications of the first ones. We will say that this expansions come from Theories (=th) on RAWs. It becomes thus clearer in what sense SeIs introduce restrictions on the codomain of F : it is possible to specify these in many different ways, but in any case restrictions coming from SeIs must be satisfied.

The general pattern of the subsequent discussion is the following :

If
(i) Obs is modelised in a way allowing him to produce such and such data representation
(ii) An NL theory is intended to account for the kind of data in (i)

then
The semantic information of a lexicon entry (i.e. SeI in lexicon quadruplets) must be such and such in order to specify the required domain of F.

The discussion intends to be a general contribution both to the methodology of lexicon building and to the correlated fundamenatl question stated by Kamp at the beginning of the 1993 Daghstuhl seminar which was : is it in principle possible without an intermediate structure such as DRS to define the F function between strings words and the world? The present contribution presents some empirical motivation arguing for a negative answer to Kamp question, if certain type of data is accepted as interesting data any theory must account for.

Assumed SOS on the Obs

The same that the ones in D-T N° 1 *plus* :

•Obs has elicited that Users consider some pairs of NL expressions as having identity of interpretation (IdI) corresponding to the same truth values and same truth conditions.

•Obs can apply +/- IdI on any pair < p, q> of NL expressions.

Observe that at this point of the modelization of Obs, Obs does not know how to describe particular truth conditions. The knowledge that is allocated to him is much more rudimentary: it consists in the possibility of ranging pairs of situations in two exclussive classes.

With this new skill allocated to Obs, Class II of Partition N° 1 can be partitioned in the following way and Partition N° 2 obtained.

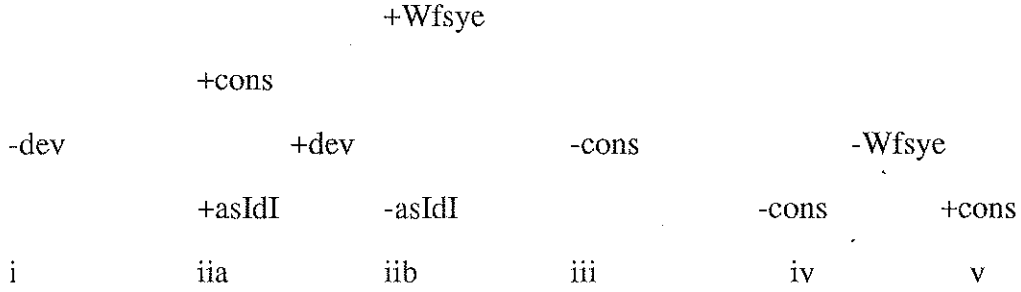
IIa A NL expression p belongs to class IIa if there is some q such that
+IdI < p, q> , and
p is +Wfsye, +cons, +dev
q is +Wfsye, +cons, -dev
Notation : +as IdI

IIb A NL expression p belongs to class IIb if for all q
+IdI < p, q> , and
p is +Wfsye, +cons, +dev
q is +Wfsye, +cons, +dev
Notation : -as IdI

+/- as IdI are thus properties of NL expressions. Intuitively, +as IdI means that there is some other linguistic way (i.e. some other expression in the NL) for saying the same thing which is not deviant; -as IdI means that there is no way to say the same thing with no deviancy

Partition N°2 on NL expressions

Classement tree



Partition N° 2 is defined on the previous classement tree

A = i U iib U iii
B = iia U iv U v

Examples of A = i U iib U iii

La fille aime le fromage
La table aime la cire
La table aime le fromage
Colorless green ideas sleep furiously

Examples of B = iia U iv
The formulations in italics illustrate 'corresponding' expressions that do not violate SeI requirements contrary to what happens with *conduire* et *essaim*..

Jacques conduit le colis à la gare
(Jacques apporte le colis à la gare)
Un essaim de chiens courait dans la campagne
(Une meute de chiens courait dans la campagne)
Un un un deux à
Un fille regarde le fromage

D-T N°2

(i) Given :

i • paradigm : as above
• data : partition N° 2
• sentence space

ii • (a-1) theory representations specified by G (a-1) map bijectively with data representations in Partition N°2

(a-2)
(F(th-r) = F(thr')) ---> for all <p, q> such that
+IdI<p, q>
< th-r, < dr, p >>
< th-r', < dr', p >>

(b) Given G and a repair function RF such that
the domain and codomain of RF are
exclusively defined in syntactic terms
For any <th-r, p[-Wfsye, +cons] > violating justified syntactic
requirements (i.e.anyone of those required by class I of the
classification tree of partition N°2, see above)
RF(<th-r, p[-Wfsye, +cons] >) = <th-r', q[+Wfsye, +cons] >
such that +IdI<p,q >

The (b) formulation of (ii) is there to clear up the issue of syntactic deviancy of sentences that are, despite their deviancy, constructible (The *Un fille regarde le garçon* example). The (b) formulation assumes that there is a RF function which obtains *Une f ille regarde le garçon.* Within this pattern the same G accounting for (a-1) accounts also for (b). This means that with respect to SeI in lexicon entries -the central issue of this contribution- nothing new must be added. Observe that RF is not intended to apply to the *Un un un deux a* example because it is not constructible.

Intuitively, the (a-2) formulation requires that anything which is different for Obs, must be specified as being associated to different denotations by <G,F > but it leaves open the possibility of a <G,F > specifying different denotations for +Id<p, q >. The (a-2) requirement cannot be satisfied by an empty SeI. But it can be satisfied by a very rudimentary

SeI, the one incorporating the only predicate 'it is different from' and then projecting differences from GPs to corresponding SeI.

(c) Given
<G,F > and <G,F >' satisfying (a-1), (a-2) and (b)
the sets of pairs of expressions +IdI< p,q > , +IdI< p,q >' for which
<G,F > and <G,F >' associate respectively representations such that
F< th-r, <dr-p >> ≠ F< th-r', <dr'-p >>
<G,F > is more highly valued than <G,F >' if
the set of pairs +IdI< p,q > is a proper subset of the set with pairs +IdI< p,q >'

Formulation (c) intends to define the mechanism for improving lexica in a principled way. It is in this perspective that phenomena illustrated by +asIdI expressions, by pre-emptions phenomena and by diathesis can be understood. Semantic features used in SeI are not selected from arbitrary set of predicates arbitrary defined but are required by improvements in the way formulation (c) is satisfied.

3.3 Specification formalism

3.4 Implementation formalism

Chapter 4

THE LEXICON AND THE CORPUS

The notion of corpus must not be mechanically equated with that of data even if corpora under convenient conditions of observation, furnish some kind of important data.

Conditions of observation of corpora must be stated. They are dependent on the adopted paradigm and on the Obs modelization.

Data extracted from corpora must be considered also in terms of negative and positive evidence in psycholinguistics studies of acquisition of natural languages.

Chapter 5

THE LEXICON AND THE WORLD

SeI in lexicon entries and thi on RAWs are related to definition, terminology and encyclopaedia.

Chapter 6

PERSPECTIVE AND EVALUATIVE PROPOSITIONS

Perspective propositions. Data-theory propositions are related to conspicuous writings in the field. Special attention is given to the possibility of adapting them even if some points in the adopted paradigm differ. For example, some semantic formal theories (dynamic semantics) do not work with truth conditions but with consistency conditions on informational states.

With respect to the frame semantics paradigm, D-T propositions differ in the following points :

• neither feelings nor intuitions are allocated to Obs. Obs is explicitly recognized and modelized.

• semantic features are not selected from some arbitrary a-priori collection of predicates coming from common sense knowledge but introduced in a principled way from general and constructive requirements.

• the introduced semantic features are required in syntactic spaces and their functionality are specified in terms of these : no semantic feature without specification of the syntactic space which requires it.

• the modelization of Obs allows to build a bridge between pre-Montague 'conceptual' semantics and truth conditional semantics.

Evaluative propositions. Specification of reflexive knowledge on lexicon KBs. Possibility of measures of fiability of encoded information.

REFERENCES

ANICK Peter, PUSTEJOVSKY James, *An application of Lexical Semantics to Knowledge Acquistion from Corpora, EUROTRA - 7 Study*

APRESYAN Yu D., MEL'CUK Igor A., ZOLKOVSKY A.K., *Semantics and lexicography : towards a new type of unilingual dictionary* . Dans Studies in Syntax and Semantics, KIEFER F. (Ed.), 1991

APRESYAN Yu D., MEL'CUK Igor A., ZOLKOVSKY A.K., *Semantics and lexicography : towards a new type of unilingual dictionary* , 1991

BES G. Gabriel, LECOMTE Alain, "Semantics feature in a generic lexicon". Dans Computational lexical semantics, Saint-Dizier Patrick, Viegas Evelyne (Eds), Cambridge, 1995

BOGURAEV Branimir, PUSTEJOVSKY James, *Lexical Ambiguity and the Role of Knowledge Representation in Lexicon Design*

CHIERCHIA Gennaro, McCONNELL-GINET Sally, *Meaning and Grammar,* An introduction to semantics, Massachusetts, 1990

CHIERCHIA Gennaro, PARTEE Barbara H., TURNER Raymond, *Properties, types and meaning ; Volume I : foundational issues,* Dordrecht Studies in Linguistics and Philosophy, 1989

CHIERCHIA Gennaro, PARTEE Barbara H., TURNER Raymond, *Properties, types and meaning ; Volume II : Semantic issues,* Dordrecht Studies in Linguistics and Philosophy, 1989

DOWTY David R., *Thematic proto-roles and argument selection* - LangageVol. 67, N°3, 1991

DOWTY David R., *Thematic roles, grammatical relations and a dynamic theory of grammar,* 1992

DOWTY David R., *Word Meaning and Montague grammar* , The semantics of verbs and times in generative semantics and in Montague's PTQ, Dordrecht Synthese Language Library, 1979

DOWTY David, R.WALL, Robert E., PETERS Stanley, *Introduction to Montague semantics,* Dordrecht Synthese Language Library, 1981

FILLMORE Charles J., Types of lexical information. Dans Langage II, 1966

FODOR Janet Dean, Semantics - Theory of meaning in generative grammar, *The language and thought series,* Harvard University Press, 1980

JACKENDOFF Ray S., *Grammatical relations and functional structure*

JACKENDOFF Ray, *Semantics and Cognition* , Current Studies in Linguistics Series, Cambridge, 1983

JACKENDOFF Ray, *Consciousness and the computational mind,* Explorations in Cognitive Science, Cambridge, 1987

KAMP Hans, *A theory of truth and semantic representation.* Dans GROENENDIJK, JANSEN and STOKHOF (eds.), Truth, interpretation and information, GRASS, Foris Publication, 1984

KAMP Hans, *Lexical meaning and conceptual structure*, Copenhagen Buisiness School, ESSLLI'94, 1994

KAMP Hans, REYLE Uwe, *From discourse to logic - Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory (Part 1),* Studies in Linguistics and Philosophy, 1993

KAMP Hans, REYLE Uwe, *From discourse to logic - Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory (Part 2),* Studies in Linguistics and Philosophy, 1993

KAMP Hans, ROSSDEUTSCHER Antje, *DRS-construction and lexicality driven inference,* Copenhagen Buisiness School, ESSLLI'94, 1994

KAMP Hans, ROSSDEUTSCHER Antje, *Remarks on lexical structure and DRS construction,* 1994

KATZ Jerrold J., Language other abstract object, New Jersey, Rowman and littlefield, 1981

KAYSER Daniel, *What kind of thing is a concept ?,* 1988

LAKOFF George, Women, fire and dangerous things - What categories reveal about the mind, Chicago, 1987

LEHRER Adrienne, KITTAY Eva F., Frames, field and constrasts : new essays in semantic and lexical organization. New Jersey, Laurence Erlabaum Associates, 1992

LEVIN Beth, PINKER Steven (Eds), Lexical and conceptual semantics, Blackwell, 1992

MEL'CUK Igor A., *Dictionnaire explicatif et combinatoire du français contemporain,* Recherches lexico-sémantiques II , Montréal

MONTAGUE Richard, Formal philosophy, New Haven, 1974

PUSTEJOVSKY James, BERGLER Sabine, *Lexical semantics and knowledge representation, Lecture Notes in Artificial Intelligence, 1991*

PUSTEJOVSKY James, BERGLER Sabine, *Proceedings of the Workshop Lexical semantics and knowledge representation,* 17 juin 1991 - Berkeley, California*1991*

PUSTEJOVSKY James, *Semantic function and lexical representation*

PUSTEJOVSKY James, *Semantics and the lexicon Studies in Linguistics and Philosophy, 1993*

SAGER Juan C., A practical course in terminology processing, John Benjamins publishing company, 1990

PUSTEJOVSKY James, *The generative lexicon,* Computational Linguistics, Vol. 17, N° 4, 1991

SOWA John F,."Principles of semantic networks", Explorations in the representation of knowledge, 1991

SOWA John F., Conceptual structures, Information processing in mind and machine,The Systems Programming Series, Addison Wesley Publishing Company, 1984

WIERZBICKA Anna, English speech act verbs - A semantic dictionary, Australie, Academic Press, 1987

WIERZBICKA Anna, The semantics of grammar, Studies in Language Companion Series, N° 18, John Benjamins Publishing Company, 1988