A GENERIC MODEL FOR REUSABLE LEXICONS:

THE GENELEX PROJECT

Marie-Hélène ANTONI-LAY
IBM FRANCE
54, rue Roger Salengro
F 94126 Fontenay sous Bois

Gil FRANCOPOULO
GSI ERLI
1, place des Marseillais
F 94227 Charenton le Pont

Laurence ZAYSSER
SEMA GROUP
16 rue Barbès
F 92126 Montrouge

Abstract

Genelex is a EUREKA project with the following goals: to define a generic model for lexicons; to design and develop software tools for lexicon management; to apply the model and the tools to dictionaries; to build full scale electronic dictionaries.

In this paper, we will discuss how Genelex model, as a generic model, fulfils the requirements of being "theory welcoming", and having a wide linguistic coverage. Syntax will be considered to demonstrate how, in syntax, with a meta-language based on "positions", Genelex provides the means for encoding syntactic information originating from different lexigraphic theories.

INTRODUCTION

All electronic lexicons are structured according to a formal model.

In this article[1], after a brief presentation of the Genelex project as a whole, we will present the strategy adopted by the Genelex consortium with regard to lexicons (G. Francopoulo). Then, with syntactic data as a basis for demonstration, we will explain how and in what terms we satisfy the claim for genericity (M.-H. Antoni-Lay & L. Zaysser).

THE GENELEX PROJECT (G. Francopoulo)

Genelex is a European industrial project dedicated to the establishment of a generic model for re usable electronic lexicons. The Genelex project has begun in October 1989 with a feasibility study that ended in September 1990.

The EUREKA phase has begun in October 1990 and will end in October 1994.

The Genelex consortium is composed of partners from Italy, Spain, Portugal and France. Each partner works on lexicons in the language of his own country. The partners in Italy are: Serv.Edi. (a subsidiary of Utet/Paravia), and Lexicon ; in Spain: Tecsidel and Univ. Autonoma de Barcelona; in Portugal: ILTEC; in France: GSI Erli, IBM France, ASSTRIL LADL, and Sema Group.

We have chosen to express the generic model as a Document Type Definition (DTD) using the Standard Generalized

Markup Language (SGML, ISO 8879). This allows us to specify simultaneously a conceptual model and a physical exchange format for re usable lexicons. The use of a standardised format has two distinct advantages.

1. A standard format makes it possible to exchange lexical data at a reasonable cost. Since the formats of the currently existing electronic lexicons vary so greatly extracting and merging lexical data can be highly complex operations. As a consequence, the exchange of lexical data is rather rare. By providing a standardised data exchange model, we hope to reduce the cost of lexical operations.
2. A standard format allows the specification and the development of a set of tools for a lexicographic workstation (loading, unloading, handling, updating and extraction). These tools can handle any dictionary respecting the generic format.

In the future, it should be possible to combine parsers and generators (possibly from different sources) which respect the same interface.

In the first stage, the main efforts of the Genelex project were applied to the definition of the generic model. But currently, the Genelex teams are developing software tools and applying thern to full scale electronic dictionaries. This application of the model to real world data provides useful feedback and allows us to improve constantly the model where needed.

LEXICONS

The trend in Computational Linguistics has been to build and manage lexicons of increasing size. This trend is likely to continue, for at least two reasons.

The first reason is that Natural Language applications keep on moving from research environments to the real world of practical applications. Since realworld applications invariably require larger linguistic coverage, the number of entries in electronic dictionaries inevitably increases.

The second reason lies in the tendency to insert an increasing amount of linguistic information into a lexicon. In the beginning of Computer Science, knowledge was encoded in program. instructions. In the sixties, an attempt was made to translate natural languages by word to word matching, using mainly morphological lexicons. The failure of this approach and the Bar Hillel report put such research into hibernation for ten years /Bar Hillel 60/. In the eighties, new attempts were made with an emphasis on grammars, but an engineering problem arose: how to manage a huge set of more or less interdependent rules. The recent tendency is to organise the rules independently, to call them syntactic/semantic properties, and to store this information in the lexicon. A great part of the grammatical knowledge is put in the lexicon (e. g. as in the LFG and GPSG systems). This leads to systems with fewer rules and more complex lexicons.

It is not less expensive to develop a natural language application with a lexicon built from scratch every time a lexicon is needed. The difference lies in diachrony: the advantages of building lexicons are best seen with time. It is now possible to factor out the linguistic information required by various applications, thereby reducing the cost per application. A lexicon can be seen as a linguistic database from which people building a specific application can extract the information they need.

In the development of cheaper and better natural language processing, the challenging task has now become the construction of re usable lexicons. To merge and extract the huge mass of unstructured data, or simply to have them used by parsers and generators, lexicons need to be structured according to a well defined model.

TERMINOLOGY

Before going into details, and to avoid any misunderstanding, we will specify what we mean by the key term: "generic".

A generic model is:

1. A model that is "theory welcoming" (see next chapter) and
2. A model that has a broad linguistic coverage. For instance, imagine a model able to express syntactic properties on verbs (given a certain theory) without being able to describe syntactic properties on prepositions (with this theory); this model could not be said "generic".

In the following, we will discuss the requirements a generic model must satisfy. Therefore, syntactic data will be taken as examples.

GENERIC SYNTACTIC MODEL (M. H. Antoni Lay, L . Zaysser)

"THEORY WELCOMING" MODEL

When constructing a lexicon, one task is clearly identified: the task of the lexicographer who records the relevant information for all entries.

To do this, he applies to all entries a heuristics and criteria given by a linguistic theory, evaluates acceptability and translates relevant linguistic data into a descriptive formalism.

It goes without saying that this work cannot be carried out without having previously defined:

1. a linguistic theory that defines the notions, or set of criteria, which decide what the linguistic facts are, and resolve the following problems: what is a dictionary entry, where do you attach prepositional phrases, how do you structure or determine the limits of a phrase, how do you distinguish an inner complement from a modifier if you decide to distinguish them.
2. a formalism (formal language) to represent and format data (definition of the formal objects handled, descriptive vocabulary, rules to handle and operate on formal objects, integrity and coherence rules).

For "paper dictionaries", editorial manuals (usually confidential in nature) carry out both tasks.

They define a set of tests and their mutual precedence, and specify the way to encode information (according to the school of reference which defines the notions). In this case, the encoding constraints are limited to the paper print issues requirements: traditional lexicography uses a formalism to structure the document and enable human reading (typographic requirements and page formatting). It is not true for Machine Readable Dictionaries (MRDs).

A human reader will easily deal with implicit references, thanks to his more or less conscious knowledge of language, lexical fields and world organisation. An MRD must collect only explicit information registered under an unambiguous format.

For machine readable dictionaries, the application usually determines the descriptive formalism and the school of reference determines the linguistic criteria to be used.

It follows that three separate skills are required to construct a lexicon: the skills of the linguist, of the "formaliser", and of the lexicographer. (These skills need not be fulfilled by three separate individuals). Every dictionary, paper or electronic, requires the use of these three skills.

GENELEX must record different kinds of descriptions in a consistent manner, taking into account that they will depend on the theoretical model, on the degree of detail and on the criteria used by the lexicographer, whatever his school of thought may be.

This will have consequences:

- at the morphological level;

- at the syntactic level;
- at the semantic level;
- on the interaction between all these levels.

It is important to distinguish between the descriptive formalism which must be theory welcoming", and the implemented dictionary which uses this formalism with reference to a certain theory. We believe that the generic approach to dictionaries can only be pursued in the area of the descriptive formalism considered as a bridge between different theories.

The GENELEX model is such a descriptive formalism. As such, and only as such, can it claim to be generic.

We want to register into this format:

- complementary but not competitive facts

  DNT (Dictionnaire Hachette de Notre Temps)
  aimer V. tr. dir.

  LADL
  aimer T7

  GENELEX

aimer	
CB	P0 PSelf (P1)
SELF	catgram: VERB trait_l: [aux:avoir]
P0	NP
PSelf	
P1	NP P[sbcat:complementizer][mood:subjunctive]

- competitive but not contradictory facts

  LADL

aimer	
Table	7
sujet Nhum	+
sujet Vl W	+
Ct ind completive pc z	+
Ct ind inf V0 W	+
Ct ind Pronom ppv le	+
Ct ind le fait que P	+
Ct ind Nhum	+
Ct ind N hum	+

IBM

```
aimer (VERB IF TRAN (INFV TR) (COPLOBJ (A P)) AUXA (QUEOBJ SUBJ) INFA INFD)
```

GENELEX

aimer[2]	
CB[3]	P0 PSelf (P1)
SELF	catgram: VERB trait_l: [aux:avoir]
P0	NP PRONOUN[lex:quelqu'un]
PSelf	V[aux:avoir]
P1	NP S[introd:le fait que] S[mood:infinitive] S[mood:infinitive][prep:à] S[mood:infinitive][prep:de] S[sbcat:complementizer][mood:subjunctive] PRONOUN[lex:le] PRONOUN[lex:quelqu'un] PRONOUN[lex:quelque chose]

To simplify, one can say that dictionaries treat similar facts in different terms. Theoretical conflicts can be solved by the formalism, especially if we succeed in defining a metalanguage that describes all phenomena. This metalanguage in syntax is based on the notion of POSITION.

POSITION AND GENERICITY IN SYNTAX

Introduction

To introduce and justify the notion of position, we will start our presentation with the category best described in the current state of syntax studies: the verb. However, the verb complementation is described in very different terms depending on the theory applied. Let's take for a demonstration:

- subject, nominal object, prepositional object, attribute: traditional grammars, in functionalism and in LFG;
- A0 ("actant"), A1, A2: verb valency;
- N0 (nominal paradigm), N1, Prep N2: distributionalism and in transformational grammars;
- NP: categorial grammars and TAG's;
- Theta roles: generative grammar;
- Arg0 (argument), Argl, Arg2: predicative logic.

We call position a meta notion that subsumes all of these.

"A position is a paradigm that gathers the various syntactic realisations of verb, subject or complement and is part of the maximal definition of verb valency." /Lay Zaysser/

J. C. Milner introduces the explicit notion of position in a more complex system that also embeds spots and places / Milner/. There is not a complete overlap with the notion introduced by Milner: we maintain the distinction between

position and position occupant, since we consider that the category of the X element and the category label of the position Y occupied by X represent independent pieces of information. We consider, lit justified to assign different labels to a position and its occupant since treating them as homonyms can create confusion.

In GENELEX, a position is represented by the symbol Pi, and defined (including its function) as the group of syntagma which can instantiate it. This symbol (Pi) makes it unnecessary to mark out a representative syntagm in the set of syntagma (marking out is necessary in Milner's theory). As a matter of fact, it is not always possible to mark out a representative syntagm.

Though, it is commonly acknowledged that NP is the prototypic subject

NP
P[sbcat:complementizer]
P[mood: infinitive]

*Example: Cette décision **regarde** Marie.*
(this decision concerns Marie)
*Qu'il prenne cette décision **regarde** Marie.*
(the fact that he takes this decision concerns Marie)
*Prendre cette décision **concerne** Marie.*
(Taking this decision concerns Marie)

It is sometimes impossible (or arbitrary) to be so positive. Let's mention:

- verbs whose complement is either direct (NP) or indirect (PP);

  Example:
  *Jean **fouille** ses poches. /Boons Guillet Leclere/*
  (Jean is going through his pockets)

- verbs of "dicere" whose objects can be: complementizers, infinitives, NPs or PPs.

  Example:
  *Je **pense** que Marie est partie.*
  (I think that Marie has gone)
  *Je **pense** à écrire*
  I'm thinking of writing)
  *Je **pense** écrire à Marie.*
  (I 'in thinking of writing to Marie)

  *Jean **dit** que Marie est partie.*
  (Jean says that Marie has gone)
  *Jean **dit** partir avec Marie.*
  (Jean says he would leave with Marie)
  *Jean **dit** de partir à Marie.*
  (Jean tells Mary to leave)
  *Jean **dit** des choses à Marie.*
  (Jean tells things to Marie)

We make the subject depend on the verb, and similarly make other determiners depend on the determined. Indeed, it is commonly acknowledged that the verb can constrain the subject.

Since the subject may have different kinds of realizations, some verbs select a subset of them.

Example:
Pierre opine de la tête. (NP subject only)
(Pierre is nodding assent)

Other verbs restrict some realizations.

Example of meteorogical verbs (impersonal subject only)
*Il **pleut**.*
(it is raining)

Example of verbs used as modals
*Il **semble** qu'il neige.*
(It seems to be snowing)
P0: NP/PRONOUN[lex:il][sbcat: impersonal]

Example of verbs taking plural subjects
*Les fourmis **grouillent**[4] dans le jardin*
(The garden is swarming with ants)
P0: NP [number:plural]

Similarly, some nouns require a definite determiner.

Example:
*La **majorité***
(most of them)
*Le **paléothique***
(the Paleolithic)

2.2 RELEVANCE OF POSITIONS TO ALL CATEGORIES

Furthermore, this definition of complementation pattern in terms of position can be extended to all categories.

Nouns

It applies to different kinds of nouns. Studies on complernentation of substantives are generally limited to verbal nominalization. The positions observed then are very similar to the verb's position.

Example:
*La terrible **destruction** de la ville par les ennemis*
(the dreadful destruction of the town by the enemies)

P0	Det
P1	AP
PSelf	
P2	PP[prep:de]
P3	PP[prep:par]

But this description also fits noun quantifiers.

Example:
*Un petit **sac** de clous*
(a small bag of nails)
*Un petit **sac** de farine*
(a small bag of wheat)

P0	Det
P1	AP
PSelf	
P2	PP[prep:de]

Some quantifiers constrain a definite determiner.

Example:
*La **plupart** des gens.*
(most people)

P0	Det[sbcat:definite]
PSelf	
P1	PP[prep:de]

Others constrain a plural complement.

Example:
*Une énorme **meute** de loups.*
(a huge pack of wolves.)

P0	Det
P1	AP
PSelf	
P2	PP[prep:de][number:plural]

The increasing specifity of phrases occupying a position permits an even more detailed subcategorization of substantives.

Adjectives

This applies also to adjectives. On the position criteria, adjectives which govern a complement will be separated from the others in the first step.

Example:
*Très **capricieux***
(very wayward)

P0	ADVP
PSelf	

*Tout à fait **légitime** que Jean râle.*
(definitely well founded for Jean to grouse)

P0	ADVP
PSelf	
P1	S[sbcat:complementizer]

Restrictions on phrases occupying a position (see section on this point below) allow even more detailed subcategorizations again.

Example:
***Content** de venir.*
(glad to come)
***Content** de ses résultats.*
(happy with his results)

P0	ADVP
PSelf	
P1	S[mood:infinitive][prep:de] PP[prep:de]

***Difficile** à satisfaire.*
(difficult to satisfy)

P0	ADVP
PSelf	
P1	S[mood:infinitive][prep:à]

Prepositions

This model can handle prepositions and conceivably other categories (conjunctions, determiners, adverbs).

Example:
***A** Paris.*
(in Paris)
***A** ma mère.*
(to my mother)
***A** boutons.*
(with buttons)
***A** faire des bêtises.*
(to play the fool)

P0	NP S[mood:infinitive]
PSelf	

2.3 POSITION AND COMPLEMENTATION PATTERN

According to positions, a single lexical entry can have one or more complementation patterns (called "Constructions" in GENELEX terminology). These patterns are subsets (subcategorizations) of the patterns permitted by the category (complementation of verbs, nouns, adjectives, prepositions)

Example:
voler (transitive)
*l'homme **vole** une pomme.*
(the man is stealing an apple)
cb: P0 PSelf (P1)

voler (intransitive)
*l'oiseau **vole** rapidement.*
(the bird flies fast)
cb: P0 PSelf

arriver (motion)
*Il est **arrivé** à Paris.*
(he has arrived in Paris)
cb: P0 PSelf (P1)

P1 is optional and defined by a subset of prepositions: locative prepositions.

arriver (modal)
*Il est **arrivé** à partir.*
(he has managed to leave)
cb: P0 PSelf P1

Here P1 is required and defined by the distribution: PP[prep:à].

The syntactic pattern for a lexical unit can be understood as an abstraction of the syntagm of which it is the head (if S V'"[5], and only for major categories) .

NOTION OF POSITION

Linearity

The term "position" is not to be understood as "place" (distinction established by Milner). As a matter of fact, a constraint on surface linear order has been drawn apart for different reasons.

1. The realization of these positions may be optional (obliteration of position). For example, the structure: P0 PSelf ((P1) (P2)) can have different realizations:
    - P0 PSelf P1 P2,
    - P0 PSelf P1,
    - P0 PSelf P2, and
    - P0 PSelf.

    Example:
    *Quelqu'un **parle** de quelque chose à quelqu'un.*
    (someone talks about something to someone)
    *Quelqu'un **parle** à quelqu'un.*
    (someone talks to someone)
    *Quelqu'un **parle** de quelque chose*
    (someone talks about something)
    *Quelqu'un **parle***
    (someone talks)

The omission of P1 should not force the renaming of P2 into P1. Positions are referred to by their name in the maximal structure.

2. Some pronominalization phenomena can reverse the order of realized positions.

   Example
   *Je **pense** à Pierre*
   (I think of Pierre)
   *J'y **pense**.*
   *Je **pense** à lui.*
   (I think of him)

3. In French and other languages, the order of the different parts of an utterance is relatively free:

   Example
   *Je **promets** de venir à Pierre*
   (I promise that I will come to Pierre)
   *Je **promets** à Pierre de venir.*
   (I promise Pierre that I will come)

A canonical linearity has been conventionally defined to supply coherent records of descriptions. The order of this canonical linearity is handled by grammar rules, or stylistic rules, that are ultimately bound to the type of position and that do not have their place in the description of construction criteria for a syntactic unit. Some lexical units do not obey these general rules. Their attachment to a certain place must be specified.

Optionality of positions

The realization of a position may be optional or required. The optionality is figured by brackets.

Example
*le roi **stipule** que...*
(the king stipulates that...)
le roi **stipule*
*(*the king stipulates)*
cb: P0 SELFPl

*Il **mange** du pain*
(He eats bread)
*Il **mange***
(he eats)
cb: P0 PSelf (P1)

The realization of one of these positions can be determined by another one. Hence, in some cases, the indirect object cannot be realized if the direct object is not.

Example
*Il **achète** un jouet à Jacques*
(He buys a toy for Jacques)
Il **achète à Jacques*[6]
*(*He buys for Jacques)*

Single or embedded brackets indicate optionality and constraints on optionality. The interpretation rule is: it is permitted to eliminate anything in brackets, but only one elimination is permitted /Lay Zaysser/

The following are examples of the application of this rule.

Given the positions P0 P1 P2, where P0 is required, we want to encode the following constraints upon optionality:

1. P1 and P2 are optional, without constraints. P0 PSelf ((P1) (P2)) is the synthetic notation for all the following realizations: P0 PSelf P1 P2, P0 PSelf P1, P0 PSelf P2, P0 PSelf.

   Example:
   *Pierre **parle** de sa soirée à Marie.*
   (Pierre talks about his party to Marie)
   *Pierre **parle** de sa soirée.*
   (Pierre talks about his party)
   *Pierre **parle** à Marie.*
   (Pierre talks to Marie)
   *Pierre **parle**.*
   (Pierre talks)
   CB: P0 PSelf ((P1) (P2))

2. P1 and P2 are optional, but if P1 is omitted, then P2 is omitted too. P0 PSelf (P1 (P2))is the synthetic notation for all the following realizations: P0 PSelf P1 P2, P0 PSelf P1, P0 PSelf.

   Example:
   *Il **achète** du pain à la boulangère.*
   (He buys bread from the baker)
   *Il **achète** du pain.*
   (He buys bread)
   *Il **achète**.*
   (He buys)
   cb: P0 PSelf (P1 (P2))

PHRASES OCCUPYING A POSITION

List

Phrases occupying a position define the distribution of a position. Phrases are structured objects. We will a priori admit a finite[7] list of symbols for phrases: S, NP, VP, ADVP, PP, AP, NOUN, PRONOUN, ADJECTIVE, ADVERB, VERB, PREPOSITION, DETERMINER, CONJUNCTION, PARTICLE, INTERJECTION, E[8]. We assume that these symbols are well known and described in an extemal grammar.

Restrictions on phrases occupying a position

One can constrain phrases by using restrictive features on components, by edicting constraints from a distance and by specifying structural properties.

1. Restrictive features on components

   The use of features (developed by unification grammars LFG, GPSG, HPSG and so on) allows the specification of different degrees of detail in a syntactic description, depending on the features actually used by the lexicographer. Features can provide a hierarchy of increasing specificity over phrases and allow pointing directly to the requested level: PP, PP[prep:à], PP[prep:à][sbcat:definite]. These features are borne by the phrase itself.

   - Morphological: Mood, Tense, Person, Gender, Number

Example:
number restriction on a component
NP[number:plural]

- Syntactic: SubCategorization, Auxiliary, Agreement, Negative

Example:
syntactic subcategorization on a component
NP[sbcat:definite]

- Lexical restriction
  - on prepositions, conjunctions, WH pronouns and any other item introducing a phrase.

    Example:
    lexical restriction on a preposition
    PP[prep:à]

  - on leaves or heads depending on whether the category is terminal or not.

    Example:
    PRONOUN[lex:il]
    NP[lex:admiration]
- Semantic: Coreference, Aspect, Class, Property Example:

restriction to animate components
NP[animate:+]

2. External constraints

In some cases, a phrase must bear further information (such as coreference) in relation with other phrases occurring in the same position or in other positions. The cb bears the feature.

- In the same position

Within the same position, it is sometimes necessary to specify that a complementizer is the result or not of a transformation from. an infinitive phrase, or that a pronoun stands for only part of the distribution of a position.

Example:
*Il **aime** dormir > il aime ça.*
(he enjoys sleeping > he enjoys that)
*il **aime** Marie > il l'aime.*
(he loves Marie > he loves her)
*il **aime** le chocolat > il l'aime / il aime ça*
(he is fond of chocolate > he is fond of it)

P1	NP S[mood: infinitive] PRONOUN[lex:ça] PRONOUN[lex:le]
Transf	(S[mood:infinitive],PRONOUN[lex:ça],pronominalization)
Transf	(NP,(PRONOUN[lex:le],PRONOUN[lex:çal]),pronominalization)

- With phrases under another position

  In some cases, it is necessary to say that the realization of a position can constrain the realization of other positions.

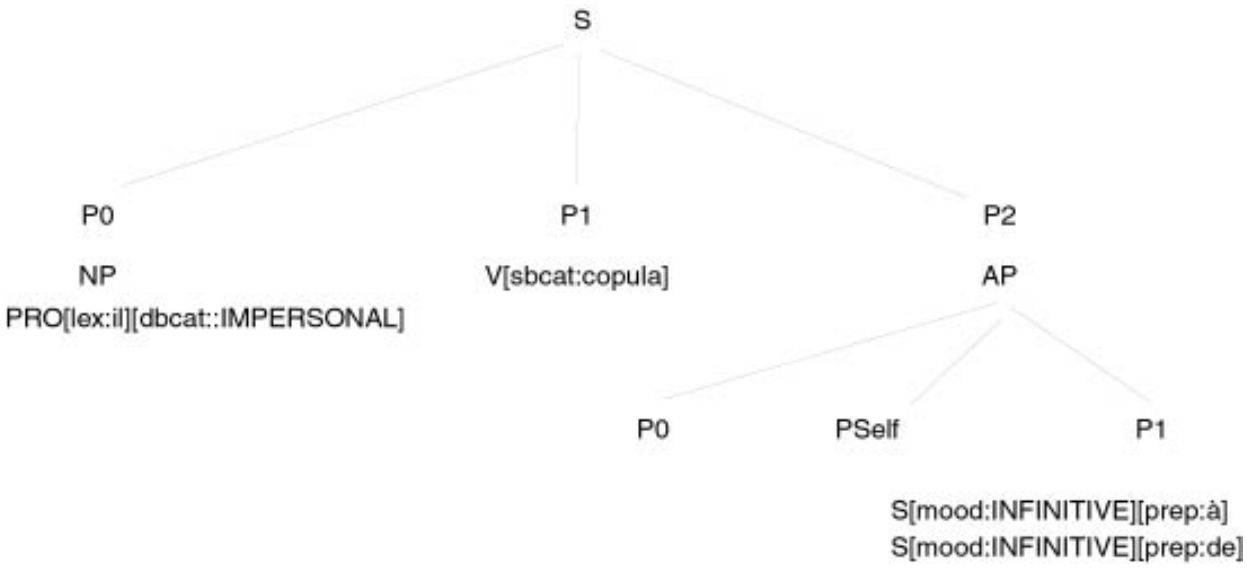
  Example:
  *Pierre **répond** à la question*
  (Pierre answers the question)
  *Pierre **répond** que c'est vrai*
  (Pierre answers that it is true)
  *Qu'il ait une telle attitude **répond** à la question*
  (His acting this way answers the question)
  Qu'il ait une telle attitude **répond que c'est vrai*
  *(*His acting this way answers that it is true)*

cb	P0 PSelf P1
P0	NP S[sbcat: complementizer][mood:subjunctive]
P1	S[sbcat:complementizer][mood: subjunctive]
cond	if P0: S[sbcat:complementizer][mood:subjunctive] Then not P1: S[sbcat:complementizer][mood: subjunctive]

3. Structural constraints

We make it possible to constrain the syntactic structure of a syntagm by describing a syntactic tree that represents it. Trees are formalised by embedded lists of positions.

Example:
Il est intéressant de travailler sur Genelex
**11 est intéressant à travailler sur Genelex*
(it is interesting to work on Genelex)
Ce document est intéressant à regarder
**Ce document est intéressant de regarder*
(This document is worth being looked at)




```
cond : if P1==PRO[lex:il][sbcat::IMPERSONAL]
       then P2.P1 = S[mood:INFINITIVE][prep:de]
       else P2.P1 = S[mood:INIFINTIVE][prep:à]
```

Genelex offers a set of tools to encode syntactic information. There is no obligation to use all of them of course, and besides, these tools accept different readings, according to the theoretical lexicographic points of view.

We assume that there are fundamentally two lexicographic trends and attempts, the "atomist" and the "syntactist".

- the "atomist" point of view is concentrated on the properties of the lexical units, is free from any grammar implication and, regarding this, only describes elements (positions) depending directly upon the lexical entry.
- the "syntactist" one is paradoxically encountered by so called "lexicalist" approaches. It registers all or part of the grammar on lexical entries and allows the description of syntactic contexts of any extent.

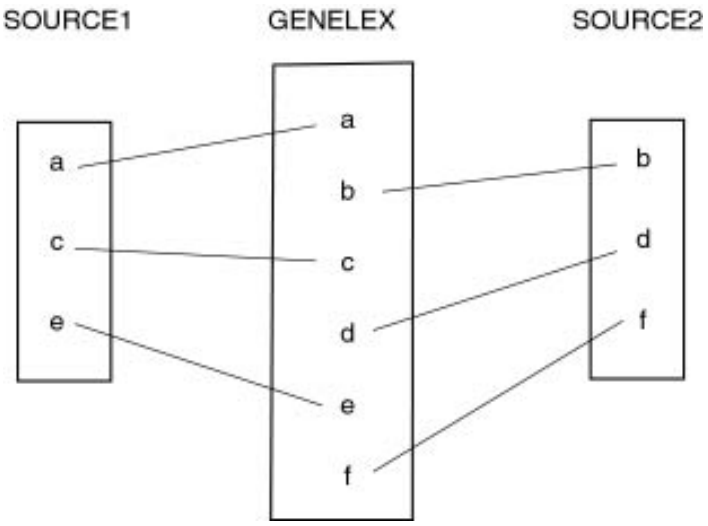
Both are taken into account by Genelex.

CONCLUSION

Genelex claims to be a generic descriptive formalism. A descriptive formalism is generic if and only if it provides the means to record linguistic facts ,displayed by various theories. However, one must distinguish several cases which require different solutions:

COMPLEMENTARY LINGUISTIC FACTS

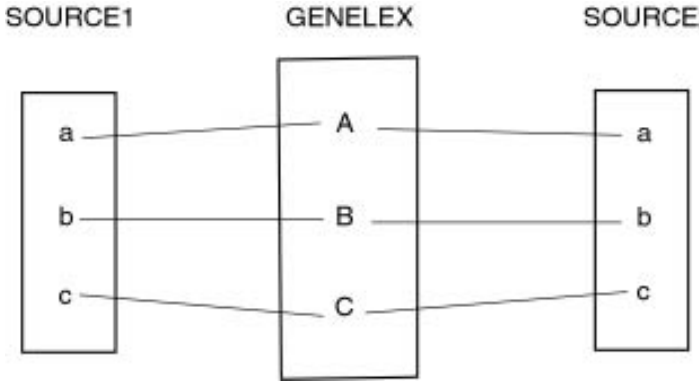
The information recorded in two dictionaries is different but complementary. In this case it is sufficient to combine the different kinds of information. Imagine a first dictionary which mentions whether a verb is transitive or not, and a second one which specifies what kind of complement is governed. Or imagine a first one that indicates graphic spelling only, while the other specifies prononciation. Both pieces of information will complete each other without any conflict.



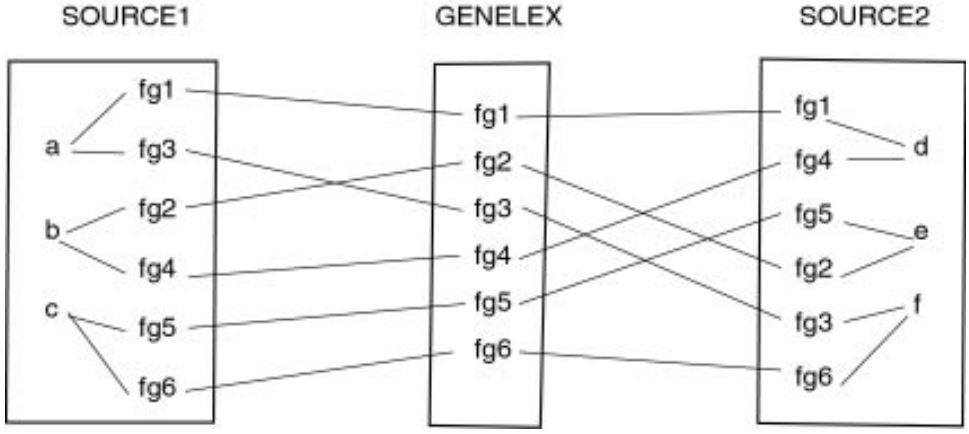
COMPETITIVE BUT NOT CONTRADICTORY FACTS

Dictionaries sometimes treat the same facts in different terms. These terminological dissensions reveal theoretical conflicts. They result more from different approaches and allowable operative definitions of linguistic objects than from the circumscribing of the objects themselves. These divergences can be solved if we succeed in:

- defining a metalanguage to describe all phenomena, that is to find equivalence classes for notions.



- identifying granular facts apart from the notion regrouping them originally, that is to find equivalence classes for granular facts entering the notions. These granular facts (elementary information) must be retrieved frorn source dictionaries and "flattened out" as a first step. Then, basic information is reorganized under the Genelex format, and is available under that format. It remains convertible into its origninal format, if necessary.



COMPETITIVE AND CONTRADICTORY FACTS

In some cases, when there is no global solution, theoretical conflicts cannot be solved. Linguistic pieces of information displayed by a theory enter clearly in contradiction with one another. Clearly, we will not allow contradictory facts to be put together in a dictionary (what about the reliability of the dictionary then ?).

Hence, for a "neutral" verb, we will not allow the same construction to be basic as well as secondary.

Example:
*Jean **casse** la branche.*
(Jean is breaking the branch)
*La branche **casse**.*
(the branch is breaking)

Same for a motion verb, one complement will not be said to be a modifer and inner complement at the same time.

Example:
*A Paris, Jean **a marché** de la place Vendôme jusqu'à la Defense.*
(In Paris, Jean walkedfrom Place Vendôme to La Defense)

Depending on the theoretical point of view, space complements can be treated differently:

- inner complements only,
- modifiers only inner complements or modifiers,
- depending on the verb.

These points of view being incompatible, one of these options must be chosen.

To guarantee genericity, one will ideally choose the theoretical option which causes the least loss of information and which permits automatic extraction of an application dictionary based on another theory.

BIBLIOGRAPHIE

ACH ACL ALLC
Guidelines for the encoding and interchange of machine readable texts.
Sperberg McQueen & Lou Burnard (eds).
TEI P1 Draft version 1. 1 Nov 1990.

G.G.Bes & K.Bashung, & A.Lecomte
Une modélisation des entrées lexicales.
Projet EUREKA GENELEX Janv. 91

G.G.Bes & M.Emorine
Une lecture des tables du Ladl en vue de la définition de la couche syntaxique de Genelex.
Projet EUREKA GENELEX Janv. 91

Y. Bar Hillel
The present status of automatic translation of languages.
in F.Alt: Advances in computer, Vol 1
Mc Graw Hill 1960

B.Boguraev & T.Briscoe
Computational lexicography for natural language processing.
Longman. 1989.

J P Boons & A. Guillet & C.Leclere
La structure des phrases simples en francais: constructions intransitives.
Droz 1976.

C. Fillmore The case for case. in Bach & Harms: Universals in Linguistics Theory.
Holt, Rinehart and Winston 1968 (reed 1972), p.1 90.

B.Fradin & J M Marandin
Autour de la définition: de la lexicographie à la sémantique.
Langue Francaise 43 1979, p.60 83.

Gazdar & Klein & Pullum & Sag
Generalized phrase structure grammar.
Harvard University Press. 1985

M. Gross
Méthodes en syntaxe: régime des constructions complétives.
Hermann 1975.

R. Jackendoff
X bar Syntax: a Study of Phrase Structure.
Cambridge, Massachussets: MIT Press. 1977

R.Jackendoff
Semantic Interpretation in Generative Grammar.
Cambridge: MIT Press 1972.

M-H Lay & L. Zaysser
Contrat Genelex Lot modèle: rapport sur la couche syntaxique.
Projet EUREKA GENELEX. Juin 91. [confidential]

J-C. Milner
Introduction à une science du langage.
Paris, Seuil 1989.

L. Tesnière
Eléments de syntaxe structurale.
Klincksieck 1959

Notes

1 This paper was published in Literary and Linguistic Computing 1994. Back to text.

2
à : to
aimer : to love
avoir : to have
le : it
le fait que : the fact that
quelqu'un : someone

quelque chose : something
Back to text.

3

cb : construction de base (basic structure)
PSelf: position of the described lexical unit
SELF: properties of the described lexical unit
P0, Pi: Position being part of the complementation pattern.
catgram : grammatical category
trait_l : list of features
prep : preposition introducing PP or S.
Back to text.

4

One use of a symmetrical verb, the other use being "*jardin grouille de fourmis*". Back to text.

5

"Three bars". Cf. Jackendoff 1977. Back to text.

6

This sentence could be acknowledged with the meaning "*he buys from Jacques*". Back to text.

7

This list contains the default values proposed by Genelex for a generic description of phrases. It is possible for users to custornize this list, but they will lack genericity then. Back to text.

8

E stands for empty category. Back to text.