Semantic Lexicons: the Cornerstone for Lexical Choice in Natural Language Generation

Evelyne Viegas[†] Pierrette Bouillon[§]
viegas@cs.brandeis.edu pb@divsun.unige.ch

[†]Computer Science Department, Brandeis University, Waltham, MA 02254 USA
[§]ISSCO, University of Geneva, 54 route des Acacias, CH-1227 Geneva, Switzerland

Abstract

In this paper, we address the issue of integrating semantic lexicons into NLG systems and argue that the problem of lexical choice in generation can be approached only by such an integration. We take the approach of Generative Lexicon Theory (GLT) (Pustejovsky, 1991, 1994c) which provides a system involving four levels of representation connected by a set of generative devices accounting for a compositional interpretation of words in context. We are interested in showing that we can reduce the set of collocations listed in the lexicon by introducing the notion of "semantic collocations" which can be predicted within GLT framework. We argue that the lack of semantic well-defined calculi in previous approaches, whether linguistic or conceptual, renders them unable to account for semantic collocations.

1 Introduction

Whether we talk of monolingual or multilingual generation, it is not surprising that there has been very little focus on the area of lexical choice. Lexical choice has often been side-stepped, not because it is a daunting issue, but rather because the interest in natural language generation (NLG) first focused on syntactic, morphological and discourse aspects of language. Semantic accuracy has been therefore sacrificed in the production of fluent grammatical sentences. In section 2, we highlight the issue of lexical choice, by arguing that generation systems must integrate lexical semantics and focusing on the treatment of Adjective-noun (Adj-Noun) collocations. We introduce the notion of "semantic collocations", which allows us to reduce the set of collocations which are usually listed in lexicons. In section 3, we present relevant aspects of the Generative Lexicon Theory (GLT), which, we argue, provides a better representation and interpretation of lexical information, enabling us to generate the set of possible semantic collocations in a predictive way without listing them in lexical entries. GLT is still under development from a theoretical point of view and up to now no generation system (as far as the authors are aware) has tried to integrate or implement its ideas. We propose to do so, and are currently studying its theoretical adequacy for generation with special reference to the issue of lexical choice. In section 4, we show that it is possible to calculate Adj-Noun semantic collocations (*a long book; an easy novel; a fast car*) as opposed to the type of collocations where idiosyncrasy seems to be involved (*a large coke* vs. *a big coke*). Finally, in section 5, we emphazise the adequacy of a framework such as GLT to generate the possible set of semantic collocations.

2 The Issue of Lexical Choice

There is a debate in NLG concerning the place of lexical choice in the generation process. Should lexical choice take place at the level of the "planning component" or the "realization component"? Even for generators which do not have a "traditional" two-component architecture, actions are still sequential and lexical choice takes place after some "planning".

Lexical choice relates to lexicalization in the sense of not only needing to pick up the right words or expressions but also of needing to "realize" them or lexicalize them. We would argue on one hand that lexicalization does not constitute an autonomous module within the process of generation, and on the other hand that lexical choice is not the sole prerogative of either the "planning" or the "realization" component. The reason is that a concept cannot be seen in isolation (the choice of a particular concept will trigger some other related concepts) and when lexicalized, the syntactico-semantics of the lexical item will impose some constraints on the further possible choice of concepts to be lexicalized (thus constraining the set of concepts triggered by the previous one). In other words in the process of production **a lexical choice can influence a conceptual choice and vice versa.**

Thus in terms of NLG this means that lexical choice has some influence at the level of "planning" and "re-

We would like to thank Susan Armstrong, Paul Buitelaar, Federica Busa, Dominique Estival, James Pustejovsky, Graham Russell and Scott Waterman for their helpful comments.

alization". Moreover, if we want to generate in an incremental way, it follows that a strict distinction between these two components can no longer hold, and that we must attempt either to bridge gaps between them (Meteer 1992) or to generate in a partly parallel fashion.

In this paper, we take the view of integrating lexical semantics in the design of the lexicon to be used in an NLG system, in order to perform the right lexicalizations. We define lexicalization as a complex dynamic process, by which we find the appropriate lexicalized items for utterances, in order to fulfill communicative goals. In fact, we think that we use a **backward and forward process** between concepts and lexical items and we believe that it is through **incremental (re)lexicalizations-(re)conceptualizations** that we perform well-formed linguistic realizations (Viegas, 1993).

In the following, after a brief overview of the issue of lexical choice, we focus on the treatment of collocations, which poses the problem of complex lexicalizations, and motivates the need of taking into account, in the process of lexicalizing, both several concepts and several lexical items.

2.1 Different Approaches

Roughly speaking, the issue of lexical choice has been investigated mainly along two different lines: a conceptual-based approach (mainly in the AI tradition) and a linguistic-based approach.[1]

Despite these efforts, lexical choice remains a burning issue. We agree with McKeown and Swartout (1988) when they say that: *"... a truly satisfactory theoretical approach for lexical choice has yet to be developed."* However, like some leading researchers in generation, we argue that it is of paramount importance to **first** know the kind of information that should be coded in the lexicon, which means to pay more attention to "the nature of words" (McDonald, 1988) and to have a *"real knowledge of [the] lexical semantics"*, as was pointed out by Marcus (1987):

> *"In some important sense, [the] systems have no real knowledge of lexical semantics .... They use fragments of linguistic structure which eventually have words as their frontiers, but have little or no explicit knowledge of what these words mean."*

In this article, we will not give a review of the issue of the lexical choice; it is enough to say that the lexical semantic component for lexical representation is still basically unused and that there is a need to tackle that issue if we want to give some new and promising impetus to the study on lexical choice.

2.2 The Treatment of Collocations

There is much divergence of opinion on just what the defining criteria for collocations are. One can minimally define a collocation as the distribution of an object or element in relation to other objects or elements, as dictionaries do; needless to say, apart from remaining vague, at best this does not provide any clue for finding them operationally.

There are three main approaches to the study of collocations, namely, lexicographic, statistical and linguistic: in each of these, the term collocation is used differently.

The traditional approach to collocations has been **lexicographic**. Here dictionaries provide information about what is unpredictable or idiosyncratic. Benson (1989) synthesizes Hausmann's studies on collocations (Hausmann, 1979), calling expressions such as *commit murder, compile a dictionary, inflict a wound*, etc. "fixed combinations, recurrent combinations" or "collocations". In Hausmann's terms a collocation is composed of two elements, a *base* ("Basis") and a *collocate* ("Kollokator"); the base is semantically autonomous whereas the collocate cannot be semantically interpreted in isolation. In other words the set of lexical collocates which can combine with a given base is not predictable and collocations must therefore be listed in dictionaries.

In recent years, there has been a resurgence of **statistical** approaches applied to the study of natural languages. Sinclair (1991) states that *"a word which occurs in close proximity to a word under investigation is called a collocate of it ... Collocation is the occurrence of two or more words within a short space of each other in a text"*. The problem is that with such a definition of collocations, even when improved, one identifies not only collocations but free-combining pairs frequently appearing together such as *lawyer-client; doctor-hospital*, as pointed out by Smadja (1993).

There has been no real focus on collocations from a **linguistic** perspective. The lexicon has been broadly sacrificed by both English-speaking schools and continental European schools. The scientific agenda of the former has been largely dominated by syntactic issues until recently whereas the latter was more concerned with pragmatic aspects of natural languages. The focus has been largely on grammatical collocations such as *adapt to, aim at, look for*. Lakoff (1970) distinguishes a class of expressions which cannot undergo certain operations, such as nominalization, causativization: *the problem is hard; *the hard-*

[1] Robin's report (1990) presents a good survey on "Lexical Choice in NLG". See also (Reiter, 1991) and (Nogier and Zock, 1992) for a comprehensive study of the evolution made in the field.

*ness of the problem; *the problem hardened.* Restrictions on the application of certain syntactic operations can help define collocations such as *hard problem*, for example. One specific proposal for how to treat collocations in a linguistic model is developed in Mel'čuk's work on lexical functions (Mel'čuk, 1988).

In this theory, lexical knowledge is encoded in an entry of the **Explanatory Combinatorial Dictionary**, each entry being divided into three zones: the *semantic zone* (a semantic network representing the meaning of the entry in terms of more primitive words), the *syntactic zone* (the grammatical properties of the entry) and the *lexical combinatorics zone* (containing the values of the **Lexical Functions** (LFs))[2]. LFs are central to the study of collocations and can be defined as the following : *a lexical function F is a correspondence which associates a lexical item L, called the key word of F, with a set of lexical items F(L) - the value of F* (Mel'čuk, 1988).

The LF **Magn**, for example, applies to different categories to deliver collocational values, expressing an intensity:

Magn(smoker) = heavy [smoker]
Magn(opposed) = strongly/vehemently [opposed]
Magn(large) = excessively [large]

The Mel'čukian approach is very interesting as it provides a model of production well suited for generation with its different strata and also a lot of lexical-semantic information. It suffers nevertheless from three main problems (Heylen et al., 1993). First, all the collocational information must be listed in a static way, because the theory does not provide any predictable calculus of the possible expressions which can collocate with each other semantically. Second, it is sometimes difficult to assign the right lexical functions for newly analyzed lexical items; if we take the example of assigning an LF to an Adj-Noun structure, it involves knowing something about the semantic relation which exists between adjective and noun. (Bloksma et al., 1993) state that *"It is precisely this information which in many cases proves extremely difficult to establish, simply because it is just not entirely clear what semantic processes are involved in the union of adjective and noun"*.

Finally, sometimes LFs are too general to be useful, as shown in the following examples:

$\mathbf{Magn}^{temp}$ (experience) = *lengthy*
$\mathbf{Magn}^{quant}$ (experience) = *considerable*
$\mathbf{Magn}_{consequences}$ (illness) = *serious*

In these cases, superscripts and subscripts are needed to restrict the scope of the LF: they enhance the precision of the LFs, making them sensitive to meaning aspects of the lexical items on which they operate, thus constraining overgeneration of multiple values; yet this also shows that the set of LFs described is not sufficient.

By contrast, our general thesis is that there is no single definition for what a collocation is, but rather, **collocational behavior emerges from a theory of what the range of connections and relations between lexical items can be.** We claim that much of the allegedly idiosyncratic and language-specific collocation in language is in fact predictable from a sufficiently rich theory of lexical organization. This is not to say that there is no need for specific lexical encoding of some idioms and phrases, but that there is seldom any attempt made to bridge the gap between conventional semantic selection and the peripheral phenomena of collocations and fixed expressions.

We will make the distinction between the following kinds of combinations:

Free-Combining Words: (*a big stick; a wonderful man; there is an old man at the door*)

Semantic Collocations: (*a fast car; a long book; to start a car*)

Idiosyncratic Lexical Co-occurrences: (*a heavy smoker* vs. *un grand fumeur* (French); *un grand/gros mangeur* (French) vs. *un gran/*gordo comelon* (Spanish))

Idioms: (*to kick the bucket; take advantage of*).

Formally, this takes us from purely compositional constructions of "free-combining words" to the non-compositional structures in idioms. The vast space between these two extremes can still be explained in terms of compositional principles with mechanisms from GLT such as type coercion and subselection (Pustejovsky, 1991, 1993), as we shall see below. Idiosyncrasies, of course, should be listed in the lexicon, yet we believe that we can reduce the set of what are conventionally considered idiosyncrasies by differentiating "true" idiosyncrasies (which cannot be derived or generated) from expressions which, since they are compositional in nature, behave predictably, and which we call semantic collocations.

3 Generative Lexicon Theory

The Generative Lexicon Theory (GLT) (Pustejovsky, 1991, 1994c) can be said to take advantage of both linguistic and conceptual approaches, providing a framework which arose from the integration of linguistic studies and of techniques found in AI. GLT can be briefly characterized as a system which involves four

[2] See (Iordanskaja, et al., 1991) and (Ramos et al., 1994), concerning the use of MTT and LFs in NLG respectively.

levels of representation which are connected by a set of generative devices accounting for a compositional interpretation of words in context, namely: the **argument structure** which specifies the predicate argument structure for a word and the conditions under which the variables map to syntactic expressions; the **event structure** giving the particular event types such as S (state), P (process) or T (transition); the **qualia structure** distributed among four roles FORM (formal), CONST (constitutive), TELIC and AGENT (Agentive); and the **inheritance structure** which involves two different kinds of mechanisms:

- the *fixed* inheritance mechanism, which is basically a fixed network of the traditional *isa* relationship found in AI, enriched with the different roles of the qualia structure;
- the *projective* inheritance mechanism, which can be intuitively characterized as a way of triggering semantically related concepts which define for each role the projective conclusion space (*PCS*). For instance in the *PCS* of the telic and agentive roles of *book* we will find at least the following predicates: *read, reissue, annotate, ...* and *write, print, bind, ...* (respectively)[3].

The most important of the generative devices connecting these four levels is a semantic operation called *type coercion* which "captures the semantic relatedness between syntactically distinct expressions" (Pustejovsky, 1994a). Another notion introduced is that of *lexical conceptual paradigms* (*LCPs*), as formalized in (Pustejovsky, 1994b). We will say that the aim of an LCP is to capture the conceptual regularities across languages in terms of cognitive invariants, like "physical-object", "aperture", "natural kind" and alternations such as "container/containee", etc. Moreover, the possible syntactic projections are associated with *LCPs*. For instance, one can say "I left a leaflet **in/inside** the book at the page I want you to read' as *book* is an *information-phys_obj-container* whereas for instance one cannot say "I put the book **in** the top of the table" as "the top of the table" is a *surface* and not a *container*[4].

In the following, we will focus on two basic mechanisms of GLT, which allow us to bridge the word usage gap, that is, on a scale of lexical specificity, from free-combining words to idioms. These are:

(1) Reference to the qualia structure: By giving every category the ability to make reference to specific semantic functions, we are encoding the "semantic basis" of word usage information with a lexical item. This gives rise to semantic collocations.

(2) Cospecification: This is the basic means of encoding specific usage information in the form of either coherent argument subtypes, or already lexicalized phrases, giving rise to idiosyncrasies and idioms, respectively.

4 Adjectival Semantic Collocations within GLT

4.1 The Semantics of Nominals

We illustrate these theoretical notions with some examples for nominals[5], paying particular attention to "covert-relational nominals"[6], that is, those exhibiting a logical polysemy. We only present partial entries, which however exhibit semantic information distributed among the qualia, thus allowing the prediction of semantic collocations as will be shown in 4.2. We give some realizations for **beer** and **writer** and discuss their representations[7]:

South African Breweries Ltd., or SAB, the country's largest producer of beer, was hit by a strike at seven of its 11 breweries around the country.

"I am a beer-drinker with a running problem," one hash lapel button reads.

$$\begin{bmatrix} \textbf{beer} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = \texttt{x:beverage} \end{bmatrix} \\ \text{QUALIA} = & \begin{bmatrix} \textbf{liquid-LCP} \\ \text{FORM} = \texttt{beer-liquid(x)} \\ \text{TELIC} = \texttt{drink(P,v:individual,x)} \\ \text{AGENT} = \texttt{produce(T,w:brewer,x)} \\ \end{bmatrix} \end{bmatrix}$$

Ms. Rifkind is a writer and editor living in New York.

Mr. Ferguson is an editorial writer for Scripps Howard News Service in Washington, D.C.

$$\begin{bmatrix} \textbf{writer} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = \texttt{x:author} \end{bmatrix} \\ \text{QUALIA} = & \begin{bmatrix} \textbf{human-LCP} \\ \text{FORM} = \texttt{human(x)} \\ \text{TELIC} = \texttt{write(T,x,v:text)} \end{bmatrix} \end{bmatrix}$$

[3] This issue is still unsettled in GLT. Our point however, being to show how to predict Adj-Noun semantic collocations, our discussion will not suffer from that lack.

[4] We follow Dubois and Pereita (1993) in their analysis of categorization in relation with cognition.

[5] For a broader account of the semantic interpretation of nominals, including nominalizations, see Pustejovsky and Anick (1988).

[6] We use "covert" to differentiate traditional relational nominals (such as *friend, father, cousin*), from the class of nouns which exhibit a polysemous behaviour (such as *book, door, record*).

[7] We mainly use the approach to typed feature structures as described in Carpenter (1992). We cannot develop here the way the information is inherited in the partial lexical entries presented.

The argument structure of nouns encodes arguments which are to be taken as logical parameters providing type information for lexical items as discussed in (Pustejovsky, 1994a). The predicates "drink", "produce", and "write", are the defaults we find in the qualia of **beer** and **writer** respectively. It is still possible to create the semantic space which these predicates belong to, through the projective inheritance mechanism.

In the cases of covert-relational nominals, exhibiting semantic polysemy, we argue that they have actually well-defined calculi. If we look at examples (1):

(1) a. *This book is heavy to carry around.*
(physical object)
b. *I read an angry book.* (text)
c. *This book is great!*
(text and/or physical object)

(1a) and (1b) illustrate the polysemy between the physical object and the notion of text, whereas (1c) can either refer to one or both aspects within the same sentence.

Traditional approaches, from transformational grammars to classical Montague grammars, account for this lexical ambiguity by postulating different entries per lexical item. These fail to capture the core semantics of the lexical items, leaving the *complementary*[8] senses unrelated. Following Pustejovsky (1994b) we suggest that covert relational nominals should have a relational structure, thus capturing polysemy within the lexical structure.

For the purpose of clarity we only give a partial representation of *book* below:

$$\left[\begin{array}{ll} \textbf{book} & \\ \text{ARGSTR} = & \left[\begin{array}{l}\texttt{ARG1 = x:text}\\ \texttt{ARG2 = y:paper}\end{array}\right] \\ \text{QUALIA} = & \left[\begin{array}{l} \textbf{information-phys\_obj-container-LCP} \\ \texttt{FORM = book\_hold(y,x)} \\ \texttt{TELIC = read(T,w:individual,x),} \\ \qquad\texttt{publish(T,v:publisher,y)} \\ \texttt{AGENT = write(T,u:writer,x),}\\ \qquad\texttt{print(T,z:printer,y)}\end{array}\right]\end{array}\right]$$

Briefly, this states that *book* inherits from the relational information-physical_object-container-*Lcp*, although imposing additional constraints of its own, represented here as the arguments, namely ARG1 and ARG2. Moreover, we have specified two defaults for the telic and agentive roles, each refering to one aspect of *book*, either **text** or **physical_object**. The sorts *publisher, writer, printer* are organized hierarchically with *individual* as a common super-type.

This nominal representation enables us to capture all the *complementary* nominal "polysemous" senses as expressed in the sentences: *The writer began his third book* (writing), *my sister began "The Bostonians"* (reading); *the binder finished the books for CUP* (binding), etc. The values of these qualia are typed and are accessible to semantic operations in composition with other phrases. One aspect of nominal representation to be captured with this formalism is the paradigmatic behavior of a lexical item in the syntax, and help understanding the processes involved in lexical selection tasks. In the next section, we address the issue of selection within the NP, and show the utility of having qualia structure associated with nouns and adjectives for compositional purposes, focusing on semantic collocations.

4.2 Adj-Noun Interpretation

Within the approach taken here, adjectives, depending on their types, will be able to modify not only the arguments of the argument structure of the nouns (ARGSTR), but also the arguments inside the agentive and the telic roles. As the information in the qualia is specific to the noun and as the same adjective can modify different roles, it is possible to deal with the polysemous behavior of adjectives and to provide a generative explanation of semantic collocations.

Very briefly, an adjective selects for a particular type, an event or an object. When it modifies an object, it selects for a particular semantic type (person, vehicle, information, etc.). When it takes an event, it can be restricted to a special type (process, event, transition) or role (agentive or telic). If the noun does not have in its argument structure the type required by the adjective, generative mechanisms can exploit the richness of typing of the qualia and generate the required type (Pustejovsky, 1994a), if it is available in the qualia and if common sense knowledge is respected. In this case, the adjective will only modify one part of the qualia (i.e. of the denotation) of the noun.

Consider, for example, the French adjectives *intelligent* (clever) and *triste* (sad) in (2). We give, for each example, the English literal translations (lit. tr.):

(2) a. *un homme intelligent/triste*; (lit. tr. a clever/sad man)
b. *des yeux intelligents/tristes*; (lit. tr. clever/sad eyes) → which show the cleverness/sadness of the person in question
c. *un livre intelligent/triste*; (lit. tr. clever/sad book) → book which shows the cleverness/sadness of the person who writes the book

[8] Weinreich (1964) makes the distinction between *contrastive* and *complementary* ambiguity. A noun such as *record* exhibits the former type between the readings *written statement of facts* and *gramophone record or disc*, and the latter between the complementary interpretation of *physical object* and *musical content*.

d. *un livre intelligent/triste* → book which causes the *cleverness/sadness of the person who reads it
e. *un sapin triste*; (lit. tr. a sad fir-tree)
→ *fir-tree that causes the sadness of the person who ...
f. *une voiture triste*; (lit. tr. a sad car) → *car that causes the sadness of the person who constructs it
g. *une robe triste*; (lit. tr. a sad dress) → *that causes the sadness of the person who wears it

These adjectives select for an object of type *person* (as shown in (2a)):

$$\left[\begin{array}{l}\textbf{triste}\\ \text{ARGSTR} = \left[\text{ARG1} = \boxed{1}\left[\ \text{person}\ \right]\right]\\ \text{QUALIA} = \left[\begin{array}{l}\texttt{change\_state-LCP}\\ \text{FORM} = \text{triste}(e^S, \boxed{1})\end{array}\right]\end{array}\right]$$

In (2bcd), despite the apparent violation of types, the modification is possible, because the qualia of the noun makes explicit different relations between the type *person* selected by the adjective (Type-Adj) and the noun (N), as:

- (N) is a constitutive element of (Type-Adj) (example (2b))
- the telic stipulates that (Type-Adj) uses (N) (example (2d))
- the agentive stipulates that (N) is produced by (Type-Adj) (example (2c))

It must be clear that this kind of modification is only possible if the relations are defined in the qualia. The sentence (2e), for example, is semantically difficult, as the word *sapin*, as a natural kind, has no telic or agentive roles (independently of particular contexts). The modication must also respect very general common sense knowledge: in (2e) and (2g), the readings **a book that causes the cleverness of the person who reads it* (2e) and **a dress that causes the sadness of the person who wears it* (2g) is blocked by common sense principles, like:

- cleverness cannot be communicated, unlike sadness
- there must be a direct causal link between the event expressed in the telic/agentive role and the sadness of the individual. This link does not relate in our societies sadness and wearing a particular dress or building a car.

Take now the case of *long*. This adjective, in one of its senses, modifies an event transition, whose it indicates the temporal duration, as shown in the examples (3):

(3) a. *le long voyage* (the long trip)
b. *un long livre* (a long book) → whose reading/writing is long

It will therefore receive the following entry:

$$\left[\begin{array}{l}\textbf{long}\\ \text{ARGSTR} = \left[\text{ARG1} = \left[\ Q_i = \text{P}(\boxed{1}e^T, x, y)\ \right]\right]\\ \text{QUALIA} = \left[\begin{array}{l}\texttt{dimension-LCP}\\ \text{FORM} = \text{long}(\boxed{1})\end{array}\right]\end{array}\right].$$

(3b) is therefore possible because events are defined in the qualia of the noun *livre*. Again, *un long sapin* has no event reading, because there is no event available in the qualia of the noun *sapin*.

The adjectives *ancient* and *former* are also event submodifiers, distinguished by the role they modify. *Ancient* is a relative adjective that submodifies the agentive role of the modified noun:

$$\left[\begin{array}{l}\textbf{ancient}\\ \text{ARGSTR} = \left[\text{ARG1} = \left[\ \text{AGENT} = \text{P}(\boxed{1}e, x, y)\ \right]\right]\\ \text{QUALIA} = \left[\begin{array}{l}\texttt{change\_state-LCP}\\ \text{FORM} = \text{distant\_past}(\boxed{1})\end{array}\right]\end{array}\right]$$

In this view, *ancient stories* (in example (3)) are *stories which were narrated in the past*, so:
distant_past$(e^T) \wedge$ narrate$(e^T, x,$ stories$)$

By contrast, the English adjective *former* is a property modifier and can only modify the telic role of the noun:

$$\left[\begin{array}{l}\textbf{former}\\ \text{ARGSTR} = \left[\text{ARG1} = \left[\ \text{TELIC} = \text{P}(\boxed{1}e, x, y)\ \right]\right]\\ \text{QUALIA} = \left[\begin{array}{l}\texttt{change\_state-LCP}\\ \text{FORM} = \text{past}(\boxed{1})\end{array}\right]\end{array}\right]$$

A *former architect* is *a person who performed his job in the past*[9], so:
past$(e^P) \wedge$ perform_the_job_of_architect(e^P, x)

In French, two adjectives with the same meaning *past* can modify these two roles: *ancien* and *vieux*, which will receive the following feature structure (which does not deal with the absolute sense):

$$\left[\begin{array}{l}\textbf{vieux}\\ \text{ARGSTR} = \left[\text{ARG1} = \left[\ Q_i = \text{P}(\boxed{1}e, x, y)\ \right]\right]\\ \text{QUALIA} = \left[\begin{array}{l}\texttt{change\_state-LCP}\\ \text{FORM} = \text{past}(\boxed{1})\end{array}\right]\end{array}\right]$$

[9] Let P be any predicate, from the qualia of the noun, and $< e_i >$, a set of ordered events; the semantics associated to *past* is then the following:
$\text{past}(e_1) \wedge P(e_1, x, y) \wedge \neg P(e_2, x, y) \wedge now(e_2) \wedge e_1 < e_2$.

That is not to say that these two adjectives will be ambiguous in context. We show elsewhere (Bouillon and Viegas, 1994) that the interpretation of the adjective can be influenced by the context or morphological and syntactical constraints as the place of the French adjective, the type of the determiner or the typography (hyphen or quotes).

Within this approach, semantic collocations can be therefore computed in the same way as other Adj-Noun constructions and do not need to be listed in the dictionary.

5 Perspectives for NLG

With GLT, we can generate dynamically the set of possible semantic collocations. This can be done incrementally, as we make available the set of possible choices at run-time, a set which will be constrained by the situational and/or contextual environment.

Suppose that we are generating Adj-Noun constructions from logical forms. From a structure like the following:

$$\exists y, x, e^T \ \text{livre}(y) \wedge \text{lire}(e^T, x, y) \wedge \text{long}(e^T)$$

we can generate two sentences: the non-collocational one *un livre long à lire* (lit. tr. a book long to read) and the collocational one *un long livre* (a long book), because the entries of the noun and the adjective in GL specify that this combination is possible.

In contrast, we will not be able to generate from the logical form below *une robe triste* (a sad dress) with the meaning of *a dress which makes me sad* because this NP is blocked by the common sense principles evocated in the previous section.

$$\exists y, x, e^T, e^S \quad \text{robe}(y) \quad \wedge \quad \text{porter}(e^T, x, y) \quad \wedge \quad \text{causer}(e^T, e^S) \wedge \text{triste}(e^S, x)$$

That is not to say that we can predict generatively all collocations. Take the examples of Adj-Noun collocations involving *grand* and *gros* with nouns denoting activities:

(4) a. *un grand/gros mangeur* (a big eater)
b. *un grand/gros fraudeur* (a big smuggler)
c. *un *grand/gros client* (a big client)
d. *un grand/*gros fumeur* (a heavy smoker)
e. *un grand/*gros professeur* (a great professor)

Here, *grand* and *gros* are intensifiers of the predicate in the telic. Un *grand fumeur*, for example, will receive the following interpretation :

$$\lambda x[fumeur(x) \ldots [Telic(x) = \lambda v \lambda e^P[fumer(e^P, x, v : tabac) \wedge grand(e^P)]]]$$

We can predict that *gros* is intensifier of the quantitative aspect of the predicate while *grand* will modify both qualitative (4e) and quantitative aspects (4abcd), depending on the salience of these aspects in the predicate (we can assume that a professor is generally judged by the quality of his courses, while a smoker by the quantity of the smoking). What we cannot do is to predict which adjective will be used with preference for the quantitative aspects.

To deal with this set of idiosyncratic lexical cooccurrences and idioms, we must take the concept of collocational information a step further, with a theory of cospecification. This takes advantage of linguistic, statistical and lexicographic approaches (see 2.2), but also adds the dimension of semantic typing, focusing on collocations as they relate to sortal selection.

For instance, the cospecifications associated with the predicates we find in the telic of *book*, namely *read*, has encoded sortal pairs, providing the privileged environment (or associations) for that word:

$$\left[\begin{array}{l} \textbf{read} \\ \text{COSPECS} = \left[\begin{array}{lcl} \text{COSPEC1} & = & \left[\begin{array}{l} \text{ARG1} = \texttt{individual} \\ \text{ARG2} = \texttt{information} \end{array}\right] \\ \ldots & & \end{array}\right] \end{array}\right]$$

In the cases of *grand fumeur* versus *gros mangeur*, we know that the telic of *fumeur* and *mangeur* (*fumer* and *manger*) are predicates, denoting activities of type process, on which we can apply a scale (*très peu ... beaucoup ... énormément ...*). The adjective which will express a point on the scale with a specific noun will be specified in the cospecifications (as below). In fact, both *grand* and *gros* can generally be understood, with sometimes a clear preference for one of these, depending of the term being modified. This preference is modelled as a partial ordering ($\sqsubseteq$) over a type hierarchy $< Cospec, \sqsubseteq >$, encoded in the cospecifications.

$$\left[\begin{array}{l} \textbf{mangeur} \\ \text{COSPECS} = \left[\begin{array}{lcl} \text{COSPEC1} & = & \left[\text{SCALE} = gros(e^P)\right] \\ \text{COSPEC2} & = & \left[\text{SCALE} = grand(e^P)\right] \\ \ldots & & \\ \text{RESTRICT} & = & cospec_i \sqsubseteq cospec_j, i < j \end{array}\right] \end{array}\right]$$

6 Conclusion

By working within the framework of GLT (Pustejovsky, 1994c) we can go beyond the "quarrel" between traditional and non-traditional architecture systems and still generate in an incremental way. This is due to the richness of the Generative Lexicon which allows for mechanisms to create dynamically on one hand the triggered concepts (by means of the inheritance structures) and on the other hand to make the syntactico-semantic information available in the linguistic environment of words (by means of the argument, event, qualia structures; and the LCPs). In this sense we have shown that GLT can be seen as a promising cornerstone for generating the most adequate lexical items.

References

Benson, M. (1989) The Structure of the Collocational Dictionary. In *International Journal of Lexicography.*

Bloksma, L., Heylen, D., Maxwell, K.G. (1993) Analysis of Lexical Functions. In D. Heylen (ed.) (1993).

Bouillon, P., Viegas, E. (1994) A Semi-polymorphic Approach to the Interpretation of Adjectival Constructions: A Cross-linguistic Perspective. In *Euralex 1994.*

Church, K.W., Hanks, P. (1989) Word Association Norms, Mutual Information and Lexicography. In *ACL 1989*, Vancouver.

Carpenter, B. (1992) *The Logic of Typed Feature Structures.* Cambridge: CUP.

Dubois, D., Pereita, H. (1993) Ontology for Concepts and Categories: is there any difference? A Few Empirical Data and Questions for an Ontology of Categorial Knowledge. In *Formal Ontology*, Padova, March 17-19.

Hausmann, F.J. (1979) Un dictionnaire des collocations est-il possible ? In *Travaux de Linguistique et de Littérature XVII, 1.*

Heid, U., Raab, S. (1989) Collocations in Multilingual Generation. In *EACL 1989*, Manchester.

Heylen, D. (Ed.) (1993) Collocations and the Lexicalisation of Semantic Information. In *Collocations*, technical report ET-10/75, Taaltechnologie, Utrecht.

Iordanskaja, L., Kittredge, R., Polguère, A. (1991) Lexical Selection and Paraphrase in a Meaning-text Generation Model. In C. L. Paris, W. Swartout and W. Mann (eds), *NLG in AI and Computational Linguistics.* Dordrecht: Kluwer Academic Publishers.

Lakoff, G. (1970) *Irregularities in Syntax.* New York: Holt, Rinehart and Winston, Inc.

Marcus, M. (1987) Generation Systems should Choose their Words. In *Third TINLAP.*

McDonald, D.D. (1988) On the Place of Words in the Generation Process. In *Proceedings of the 4th International Workshop on NLG.*

McKeown, K.R., Swartout, W.R. (1988) Language Generation and Explanation. In Zock, M. and Sabah, G. (Ed) *Advances in NLG.* NJ: Ablex.

Mel'čuk, I. (1988) Paraphrase et lexique dans la théorie Sens-Texte. In G. Bes, C. Fuchs (ed.) *Lexique 6.*

Mel'čuk, I., Arbatchewsky-Jumarie, L., Elnitsky, L., Lessard, A. (1991) *Dictionnaire explicatif et combinatoire du français contemporain.* Montréal : Presses de l'université de Montréal.

Meteer, M. (1992) *Expressibility and the Problem of Efficient Text Planning.* Great Britain: Pinter Publishers Ltd.

Nogier, J.-F., Zock, M. (1992) Lexical Choice by Pattern-matching. In *Knowledge Based Systems*, 5 (3).

Pustejovsky, J., Anick, P. (1988) On the Semantic Interpretation of Nominals. In *Coling 1988*, vol.2: 518-523.

Pustejovsky, J. (1991) The Generative Lexicon. In *Computational Linguistics*, 17(4).

Pustejovsky, J. (1993) Type Coercion and Lexical Selection. In J. Pustejovsky (ed.) *Semantics and the Lexicon.* Dordrecht: Kluwer Academic Press.

Pustejovsky, J. (1994a) Linguistic Constraints on Type Coercion. In P. St-Dizier and E. Viegas (Eds) *Computational Lexical Semantics.* Cambridge: CUP.

Pustejovsky, J. (1994b) Semantic Typing and Degrees of Polymorphism. In C. Martin-Vide (Ed) *Current Issues in Mathematical Linguistics.* Elsevier North Holland Inc.

Pustejovsky, J. (1994c) *The Generative Lexicon.* MIT Press.

Ramos, M., Tutin, A., Lapalme, G. (1994) Lexical Functions of Explanatory Combinatorial Dictionary for Lexicalization in Text Generation. In P. St-Dizier and E. Viegas (Ed) *Computational Lexical Semantics.* Cambridge, NY: CUP.

Reiter, E. (1991) A New Model for Lexical Choice for Nouns. In *Computational Intelligence, 7(4): Special Issue on NLG.*

Robin, J. (1990) *Lexical Choice in NLG.* Technical Report CUCS-040-90, Columbia University, New York.

Sinclair, J. (1991) *Corpus, Concordance, Collocations.* Oxford: Oxford University Press.

Smadja, F., McKeown, K. (1991) Using Collocations for Language Generation. In *Computational Intelligence, 7(4): Special Issue on NLG.*

Smadja, F. (1993) Retrieving Collocations from Texts: Xtract. In *Computational Linguistics, 19(1).*

Viegas, E. (1993) *La lexicalisation dans sa relation avec la conceptualisation: problèmes théoriques.* Doctorat Nouveau Régime, Université Toulouse le Mirail.

Wanner, L., Bateman, J. (1990) Lexical Cooccurence Relations in Text Generation. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, Cambridge, MA.

Weinreich, U. (1964) Webster's Third: A Critique of its Semantics. In *International Journal of American Linguistics* 30: 405-409.