In: ***Proceedings of the Fourth Symposium on Natural Language Processing 2000*** **(SNLP'2000). Chiangmai, Thailand, 10-12 May 2000, pp. 37-50.**

A "NATURAL" LEXICALIZATION MODEL FOR LANGUAGE GENERATION

A. POLGUÈRE
OLST, Department of Linguistics and Translation, University of Montreal, C.P. 6128, Succ. Centre-Ville, Montreal (Quebec) H3C 3J7 CANADA

We propose a general lexicalization model which accounts for how lexical units are selected and introduced in linguistic utterances during language generation. This model aims at "naturalness" by being based on actual lexical knowledge used in speech; consequently, it should be compatible with standard patterns of behavior shown by humans when they speak (flexibility in computing both content and form of linguistic utterances, prototypical types of mistakes and backtracking, etc.). The main advantage of our model, once implemented in automatic language generation, is that it takes into account fundamental differences that exist between lexical units, with regard to why and how they are used in texts. This is achieved by means of a stratificational approach to lexicalization, where each type of lexical unit is introduced at a proper level of representation, according to the role it plays in the enunciation.

Section 1 offers a general characterization of the approach and makes explicit its main assumptions. Sections 2 to 4 successively examine the three levels of transition implied by the stratificational structuring of the model. Section 5 concludes with an examination of its relevance to the design of text generation systems.

Keywords: language/text generation, lexicalization, lexical choice, Meaning-Text theory.

1. GENERAL CHARACTERIZATION OF THE PROPOSED LEXICALIZATION MODEL

1.1. Aims

Lexicalization in language generation is the process by which lexical units are selected and introduced in natural language utterances. Whereas a rather large number of recent publications deal with lexicalization in the context of text generation systems (see for instance Elhadad *et al.* 1997, Stede 1995, 1996, 1998 and Wanner 1997), we found that all text generation systems whose design we had the opportunity to study fail to achieve lexicalization in a way that reflects the very hybrid nature of this complex process when it is performed by human speakers. To take just one specific example, we do not know of any text generation system (other than, to a certain extent, the system presented in Iordanskaja *et al.* 1996) that would clearly distinguish between the selection of *big* in a context like *Paul is a big man* vs. *This is a big mistake*. While such a distinction may seem unnecessary and undesirable to many at first glance, we believe it to be fundamental from a lexicalization point of view and we will try to demonstrate this point later in the paper.

What we present here is a "*NATURAL*" *LEXICALIZATION MODEL*, i.e. a model which aims at accounting for lexicalization performed by both humans and machines. By saying *lexicalization performed by humans* we do not refer to actual cognitive processes —i.e. what takes place in people's head when they talk, but to human linguistic behavior —i.e. what people do or appear to be doing when they talk. The observation of human speech behavior is particularly useful when something goes wrong in the generation process. It is a well-known fact that mistakes, backtracking and the like often reflect problems of access to linguistic knowledge, and can help us draw up blueprints for modelling this knowledge.

Let us illustrate this point with one very concrete example. We know that often, if not most of the time, people start to utter the beginning of a sentence without knowing exactly how they are going to end it. This observation implies that a word occurring in position P in a sentence is more likely to have been computed by the speaker before a word occurring in position P+n (where $n \geq 1$).

© 2000 The Fourth Symposium on Natural Language Processing

However, in some cases, it is possible to state with almost certain accuracy that the opposite order of computation applies. Take for instance the following sentence:

(1) *A small group of soldiers staged a successful coup against the President.*

It is almost certain that, under normal circumstances, the speaker uttering (1) will have computed *coup* before he computed *staged*, even though the latter occurs in the sentence before the former. This is confirmed by the observation that, for instance, the pause in a spoken sentence such as (2a) is more likely than the pause in (2b):

(2) a. *We strongly condemn this — **er** — coup.*
b. *A small group of soldiers staged a successful — **er** — coup.*

To be more accurate, we should say that both pauses, if indeed realized, should not be interpreted in the same fashion. The pause in (2a) could arise from an exhausted (or non-native) speaker trying to retrieve the word *coup* from his failing memory, by it could also be produced because the speaker hesitates to utter this word. In (2b), on the other hand, we most certainly have a case of a hesitant speaker who does not dare to utter the word *coup*, for whatever reason. These predictions and the corresponding observations that would, hopefully, confirm them, originate from the fact that *to stage a coup* is a collocation —a very special type of linguistic expression, which we will be looking at in Section 3.1 below. It is our contention that a "natural˘ lexicalization model should handle this type of expression in a particular way, a way that reflects its linguistic specificity.

We believe that our lexicalization model is not only compatible with actual linguistic knowledge and human linguistic behavior but also that it is suitable for the design of text generation systems with strong paraphrasing capabilities. In other words, it allows for the construction of systems which, like any speaker of a given language, are capable of establishing a correspondence between a given meaning one wants to express and all (paraphrastic) sentences that express this meaning. Text generation systems should of course not necessarily produce paraphrases, but they should have the power to do so because, like any speaker, they should possess the knowledge that allows for paraphrasing to be performed. This knowledge is truly linguistic in nature and we would like to suggest that it is the key to the implementation of an efficient and flexible linguistic synthesis process.

1.2. Main assumptions

In proposing a natural lexicalization model, it is necessary to set out a series of assumptions as points of departure, with the hope that they will prove both powerful and appropriate enough to lead to an effective implementation of lexicalization in language generation. We will thus start with a brief presentation of our four main assumptions.

Assumption 1. A natural lexicalization model should be stratificational. It should not be based on a single module of "lexical choice" but rather should involve multi-faceted computation of lexical units.

This assumption makes our model quite compatible with De Smedt (1990)'s lexicalization strategy, where the computing of lexical units is dispatched throughout the realization process in interaction with "grammaticalization." This can however be contrasted with the approach taken in many traditional text generation systems, where lexicalization is being taken care of by a single, semi-autonomous module performing the so-called lexical choice. This strategy has been harshly but soundly criticized in Bateman (1998): *Trying to construct a 'module' to do this* [= lexical choice] *is poor engineering (among other things) since the complexity of the module is equivalent to the task being attacked as a whole: generating the correct word turns out not to be that much more straightforward than generating an appropriate text.*

The need for a stratificational approach to lexicalization has at least two origins, both of which will be examined in this paper: 1) deep differences in nature that exist among lexical units, and 2) differences in the choices performed during lexicalization.

Assumption 2. A natural lexicalization model should be embedded in a specific linguistic theory.

We believe that lexicalization is not that different from grammaticalization with regard to its interdependence with the actual linguistic model (grammar and lexicon) it presupposes. It is therefore essential to present a natural lexicalization model within the framework of a complete structured linguistic theory so that theoretical "prejudices" are made absolutely clear. We have adopted *MEANING-TEXT THEORY* (Mel'čuk 1981; Mel'čuk and Polguère 1987) as our theoretical framework. This means that the model we propose would have to be reinterpreted if it is to cohabit with other approaches to the modeling of natural languages. Our hope is of course that essential concepts we introduce here will survive such a reinterpretation process and serve as a source of inspiration for researchers working within alternative frameworks.

Assumption 3. The input to a linguistic model (grammar and lexicon) used in language generation is a representation of a linguistic message, i.e. an informational content destined to be linguistically expressed in a sentence.

Taking a blackbox view of a linguistic model of a given language, we can say that such a model (grammar and lexicon) is to be fed with an input that concerns only the sentence level. The fact that people actually compute and fluently generate sentence after sentence demonstrates that it is conceivable to state that lexicons and grammars are suited for handling speech production at the sentence level. Moreover, certain text generation applications, by their very nature, do require processing to be performed at the sentence (or, even, sub-sentence) level. Such is the case of so-called real-time commentary generation, beautifully demonstrated in Tanaka-Ishii *et al.* (1998).

Assumption 4. If we make use of a linguistic model that functions as a paraphrasing engine, no lexical choice is required at the level of message computing.

Messages are not made up of lexical units. They are made up of selected meanings, each of which indicates a set of possible lexical realizations. As we will show later, actual proper lexical realizations will be handled by a linguistic model provided it not only contains linguistic knowledge but also functions as a *PARAPHRASING ENGINE*: it should be able to associate to any cluster of meanings forming a message all possible paraphrases expressing this message. This is probably the main feature of the Meaning-Text approach and this is why we use it as our framework of reference in the present proposal —in line with research that we have previously been involved in (Iordanskaja *et al.* 1991).

Our theoretical choices having been made explicit, we can now proceed with the actual description of our lexicalization model proper. It is based on four levels of representation —conceptual, semantic, deep-syntactic and surface-syntactic— which imply three transitions: conceptual-semantic, semantic-deep syntactic, and deep syntactic-surface syntactic. This does not mean that no other level of representation is required for the full modelling of language generation; we only consider here levels pertinent to the lexicalization process. In the following sections we present in turn each of these transitions, together with the various types of lexicon-related choices they involve.

2. THE CONCEPTUAL-SEMANTIC TRANSITION

2.1. Characterization of the conceptual representation: modelling a state of affairs

A lexicalization model that accounts for all choices directly related to the selection of lexical units in sentences has to start with the representation of the given state of affairs the target sentence will be about: a *CONCEPTUAL REPRESENTATION*. In human speech, this state of affairs is an infinitely

complex one as it comprises all data connected to the pragmatic context of communication: what the speaker wants to talk about, what he knows, what he believes in terms of his and the addressee's knowledge, etc. On the other hand, in text generation applications, conceptual representations can easily boil down to huge databases of facts (statistical data, structured text corpora, etc.). We therefore do not want to get into the debate here on the actual formats that are appropriate for conceptual representations —whether conceptual graphs *à la* Sowa (1984) or logical formulas of a knowledge representation language. Suffice it to say that this level of representation encodes factual knowledge that will be used to compute linguistic messages. Using a well-known terminology of text generation (see McKeown 1985), we can say that a conceptual representation belongs to the *STRATEGIC* module of text generation (which performs so-called text planning), while the next level of representation —the semantic— is the actual input to the *TACTICAL* module (which performs linguistic realization/synthesis).

For the sake of illustration we will presuppose in this paper a conceptual representation of facts concerning travels done by heads of state. These facts may be described in terms of "main actors" (heads of state and their immediate collaborators), places visited, dates of visit, official and unofficial purposes of the visit, etc. For instance, a tabular representation of a sample fact could be:

```
( mainActor::US_President(Bill_Clinton),
  place::Uganda,
  date::1998,
  duration::10.days,
  purpose::official )
```

The representation we consider is conceptual in that those facts are already structured and recorded according to a given perspective on the situation in question. Other perspectives on the "real-world" facts will of course lead to different conceptual representations. The essential point to retain here is that conceptual representations represent states of affairs, not linguistic content of texts.

2.2. Characterization of the semantic representation: modelling a message content

Each sentence we produce conveys a message, which can be complex (i.e. made up of several elementary propositions). In addition, for each sentence we utter, there is often a large number of other sentences that we could have uttered in order to express more or less the same message: alternative paraphrases. We call the representation of a message a *SEMANTIC REPRESENTATION*. Such a representation contains the propositions to be conveyed —modelled as predicate-argument clusters— and a given communicative packaging of these propositions —called the *COMMUNICATIVE STRUCTURE* of the message.

Let us consider a simple event: the US president's visit to Uganda in 1998, already mentioned above. Many things may be said in just one single sentence about this visit, but we will imagine that the propositional content a speaker (or a machine) has selected to express in a sentence is simply that 1) the US president went to Uganda and 2) he did it in 1998. Furthermore, we will decide that the pragmatic (textual or other) context leads us to package this informational content in the following way: the date of the trip is presented as being the information actually communicated —the *RHEME* of the message— and this is communicated about the trip to Uganda by the US president —the

THEME. Figure 1 below is an illustration of the semantic representation of this simple linguistic message; sentences (3a-b) are two paraphrases that can be used to express this same message.

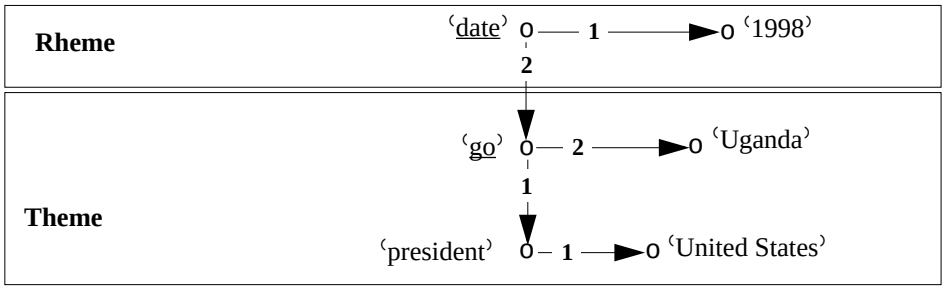

FIGURE 1. Semantic representation for (3a-b)

(3) a. *The American president went to Uganda in 1998.*
b. *The president of the United States made a trip to Uganda in 1998.*

Following Meaning-Text conventions we have encoded the semantic representation by means of a directed graph where nodes represent linguistic meanings of the target language and arcs, predicate-argument relations. The communicative packaging is represented by means of subgraphs (here, Theme and Rheme) whose dominant nodes, i.e. central meanings, are underlined. For instance, the fact that the meaning ʿgoʾ is underlined tells us that the Theme is about an event (*The American president went to Uganda…*) and not about a person (*The American president who went to Uganda…*). Notice that it is the specification of the communicative structure of the message that prevents us from associating sentences such as (4a-b) below to Figure 1. Though expressing roughly the same propositional content, these sentences possess very different communicative organization from that of (3a-b) and cannot be considered valid paraphrases for them. In other words, they do not express the same message.

(4) a. *The president who went to Uganda in 1998 is the American president.*
b. *It is the American president who went to Uganda in 1998.*

The semantic representation corresponding to (4a) is:

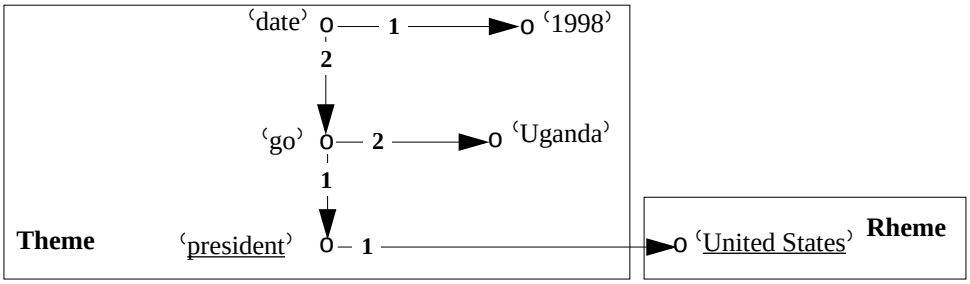

FIGURE 2. Semantic representation for (4a)

Notice that this figure involves the same set of predicate-argument relations as Figure 1, of which it differs only by its general communicative "topography." The semantic representation corresponding to (4b) is almost identical, except for the Rheme that ought to be flagged for focalization.

2.3. Lexicon-related choices performed during the conceptual transition

We emphasize that the building of a message —or the construction of a semantic representation for it— does not involve any lexical choice; at this point, only meanings are chosen. Though the choice of meanings is contingent upon whether lexical units for expressing these meanings do or do

not exist in the target language, one should see the selection of a particular meaning as only a strong option taken by the speaker (or the text generation system) on the set of possible lexical units that express this meaning. Mathematically speaking, a meaning is the set of all lexical units, or configurations of lexical units, that can express it. For instance, the meaning ‹United States› in Figures 1 and 2 does not point directly to the expression *United States*, but to the **set** of all paraphrastic expressions that will be suitable for expressing the name of this country when it functions semantically as argument of the predicate ‹president›; namely: ***American*** [*president*], ***US*** [*president*] and [*president of*] ***the United States***.

In addition, we consider that the lexical units actually used need not express the exact meanings contained in the message. For instance, there surely is a slight difference in meaning between *to go to a country* and *to make a trip to a country.* A dictionary for natural lexicalization should provide us with the indication that both expressions are quasi-synonymous, in addition to the specification of meaning differences (via the lexical definitions of GO and TRIP). An approximate and semi-formal definition of ‹X's trip to Y› could be:

(5) go(X, Y) = α
& has_purpose(α)
& short(stay(X, Y))

This formula allows us to see immediately that *X's trip to Y* is more specific —i.e. expresses more— than *the fact of X going to Y.* Therefore, if the strategic module of a text generation system inserts the meaning ‹go› in a message, it will be up the tactical module to (i) identify that the noun TRIP is a possible lexicalization of this meaning, (ii) identify that such lexicalization would entail the expression of additional (unplanned) meaning and (iii) send a request to extra-linguistic modules of the system asking them to check whether the expression of this additional meaning will be acceptable in the current context of enunciation. Of course, these operations, as well as most of the operations entailed by our lexicalization model, presuppose rather sophisticated lexicons, such as those described in Mel'čuk and Polguère (1987), Polguère (2000) or Wanner (1999). The construction of such lexicons is a prerequisite to the successful implementation of natural lexicalization models.

Throughout our presentation we suppose that any choice of a unit U —whether a meaning or a lexical unit— performed by our lexicalization model at a given transition phase can fall into one of the following three categories:

- $\rightarrow$**U** the choice of U is mainly determined by constraints originating from the source level of the transition;
- $\rightarrow$**U**$\leftarrow$ the choice of U is equally determined by constraints originating from the source and target levels of the transition;
- **U**$\leftarrow$ the choice of U is mainly determined by constraints originating from the target level of the transition.

If we apply this to the conceptual-semantic transition, we get three possible choice patterns for a given meaning ‹m›.

In a $\rightarrow$**‹m›** **choice**, the meaning is introduced because it has been determined that the corresponding information is to be communicated. This is the default situation, and we can easily imagine a context of communication where all meanings in Figure 1 would have been selected this way.

In a $\rightarrow$**‹m›**$\leftarrow$ **choice**, the meaning is introduced in order to satisfy two types of goals: firstly, for extralinguistic reasons, in order to convey a given informational content; secondly, for purely grammatical reasons, to provide the linguistic model with informational content that is required in order to produce grammatical utterances. Such can be the case with *GRAMMATICAL MEANINGS* like those expressed by tenses in English. A tense can be expressed because one wants to actually state that an event is located in the past, present or future. But it is also a meaning that one has to express with the

main verb of an English sentence. Choosing to express temporal information about a given event also provides the synthesizer with information needed by the English grammar in order to express this event by means of a verb. The situation is totally different in languages such as Mandarin, for instance, where there is no grammatical tense. In these languages, a temporal characterization of an event results exclusively from (conceptual) communicative needs.

In a **‹m›$_{\leftarrow}$ choice**, the meaning is introduced strictly for grammatical reasons. This again can be the case for grammatical meanings expressed by English tenses. Figure 1, for instance, reflects a situation where no communicative needs arose that would lead us to express whether the trip took, take or will take place —maybe because the context (we are in 2000) would suffice. However, the English grammar will not let us proceed without expressing the tense, and it will have to be computed and inserted in our message in the semantic-deep syntactic transition. We voluntarily omitted grammatical meanings in our example in order to show that, even though our model is a transitional and stratificational one, choices ought not to be performed in a strictly sequential fashion. **Due to the lexical and grammatical characteristics of the target language, each transition may force us to express something that was not initially planned.**

3. THE SEMANTIC-DEEP SYNTACTIC TRANSITION

3.1. Characterization of the deep-syntactic representation: a tree hierarchy of deep lexical units

If we try to establish a correspondence between Figure 1 and sentence (3b) we can see that all words used in this sentence do not have the same role to play relative to the expression of the message. Only a few lexical units can be directly traced back to the propositional content of the message. We call these units *DEEP LEXICAL UNITS*. They possess three main characteristics:

1. they are very numerous and form the lexical core of any natural language;
2. their meaning can be described with an analytical definition —e.g., ‹X goes to Y› = ‹X makes Y be his new location …›;
3. because they possess "true" lexical meaning, they are the lexical units that really matter from the point of view of the speaker's communicative goals and therefore from the point of view of language generation.

We identify six deep lexical units in (3b), namely: PRESIDENT, UNITED STATES, TRIP, UGANDA, IN and 1998. One can see that this list contains both open-class words (four nouns and one numeral) and a closed-class word (the preposition IN). This shows that, contrary to a very commonly accepted viewpoint, we do not believe the open- *vs.* closed-class words distinction to be the most relevant one for lexicalization. For us, it does not matter so much whether a lexical unit belongs to an open- or closed-class. What truly matters is what this unit expresses and, as we will see right now, whether its selection is linguistically contingent upon the selection of other lexical units. All six units listed here have the above-mentioned characteristics of deep lexical units.

Now, let us move to a delicate question. MAKE (in *made a trip*) could at first sight be considered as meeting these requirements. However, the use of *to make* in order to express something like ‹to accomplish (an action)› is contingent upon the word chosen to express the action in question. For instance, we say *to make a trip* but *to pay a visit* (see $^{?*}$*to make a visit* and **to pay a trip*). We are faced here with a clear case of *COLLOCATION* in the Meaning-Text sense (Mel'čuk 1995). To put it briefly, a *COLLOCATION* is a semi-idiomatic expression made up of a *BASE*, e.g. *visit*, chosen by the speaker strictly on the basis of its meaning, and a *COLLOCATE*, e.g. *to pay* in *to pay a visit*, chosen by the speaker in order to express a very general meaning (here, something as vague as ‹to do›) contingent upon the choice of the base. Following Meaning-Text modelling of collocations we consider that the entity of the deep-syntactic representation corresponding to *to make* is not the verb MAKE itself but the application of the *LEXICAL FUNCTION* **Oper$_1$** to the unit TRIP: **Oper$_1$**(TRIP). **Oper$_1$**(L)

denotes the set of "empty" verbs (also called *support verbs*) that take the nominal lexical unit L as first complement and the expression of L's first semantic actant as grammatical subject —on lexical functions and their implementation, see Wanner (1996). It is important to notice that the application of a lexical function LF to a given lexical unit L —i.e. LF(L)— points to a **set** of possible realizations and not just to one; for instance:

Oper$_1$(TRIP) = *to take*, *to make* [ART ~], *to go* [*on* ART ~].

Oper$_1$(L), which has a virtually empty semantic content, is an extreme case of LF application. In many cases of collocations, collocates do express meanings that are important at the message level. Such is the case of intensifiers —***driving***/***heavy***/***pouring***/***soaking***/***strong***/***torrential*** *rain*, *to regret* ***deeply***/***bitterly***/***very much***, verbs of "realization" —***to carry out***/***execute***/***follow*** *an order*, ***to achieve*** *a dream*, and so on. In other words, the entity LF(L) seems to function pretty much like a deep lexical unit as far as the expression of meanings is concerned. We therefore consider that applications of lexical functions (but not values/collocates themselves!) ought to be introduced at the same level as "standard" deep lexical units.

Figure 3 below gives the *DEEP-SYNTACTIC REPRESENTATION* for sentence (3b) and for all its paraphrases that would be made up of the same deep lexical units and LF applications, organized in the same general syntactic structure; examples of such paraphrases are given in (6a-b).

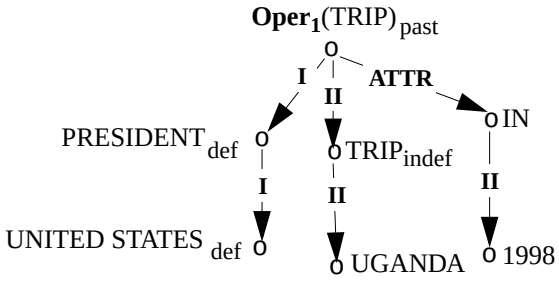

FIGURE 3. Deep-syntactic representation for (3b, 6a-b)

(6) a. *The president of the United States made a trip to Uganda in 1998.*
b. *The American president went on a trip to Uganda in 1998.*

Note that, in accordance with the Meaning-Text approach to grammar, we use *DEPENDENCY TREES* (Tesnière 1959; Mel'čuk 1988) to express the sentence syntactic organization.

In order to produce Figure 3 from the message in Figure 1, "missing" information about grammatical tense has been computed, together with other necessary grammatical information, such as the definite/indefinite marking of PRESIDENT, TRIP and UNITED STATES. We cannot delve here into details of the formalism and of how such representation is structured. Suffice it to say that all transitions have to be performed under the direct control of the grammar and the lexicon of the language in question, with possible requests to extra-linguistic knowledge. For instance, the grammar rule that will associate the deep syntactic node PRESIDENT$_{def}$, in Figure 3, to the semantic node ‹president›, in Figure 1, will have to (i) identify that information on definiteness is required for the targeted construction to be grammatical, (ii) look for this information in the message, (iii) call for this information to be computed by extra-linguistic modules of the generator once it fails to find it in the message. The strategy described here is quite similar to the example given above (Section 2.3) for lexical choice (*to go* vs. [*to make*] *a trip*). It is yet another illustration of the fact that non-sequential computing of semantic content of utterances ought to be implemented in order to achieve a natural lexicalization strategy.

3.2. Choices performed during the semantic-deep syntactic transition

As both deep lexical units, DL, and applications of lexical functions, LF(DL), can be introduced at this level of transition, we get six logical types of choices to be handled by our lexicalization model during the semantic-deep syntactic transition: $_{\rightarrow}$DL, $_{\rightarrow}$DL$_{\leftarrow}$, DL$_{\leftarrow}$, $_{\rightarrow}$LF(DL), $_{\rightarrow}$LF(DL)$_{\leftarrow}$, and LF(DL)$_{\leftarrow}$.

In a **$_{\rightarrow}$DL choice**, a DL is selected and introduced in the deep-syntactic tree solely to express part of the message. This is a very rare case as grammatical considerations normally play an important role in the selection of DLs (see next case below). However, this type of choice could be performed to express logical connectors, for example, for which the speaker has very little freedom (see for instance the choice of *and* in *Bill and Hillary travelled*).

In a **$_{\rightarrow}$DL$_{\leftarrow}$ choice**, a DL is selected both in order to express a given meaning as well as to meet some grammatical requirements: it has to find its place in the deep-syntactic tree. This is the prototypical case of DL selection. For instance, the selection of PRESIDENT meets two requirements: it allows us to express the desired meaning as well as to build a correct nominal subject for the main verb. Note that we have more options in the expression of the meaning ‹United States› as a syntactic dependent of PRESIDENT: its realization can be either nominal —*the president of the United States*— or adjectival —*the American president*.

In a **DL$_{\leftarrow}$ choice**, a DL is selected only in order to "hold" part of the deep-syntactic tree. Such a DL is often very special in nature and does not really have the second characteristic of normal DLs: it does not possess a fully definable meaning. It is the case, for instance, of *times* in *He went four times to Brazil*, which is used only to support some form of quantification of the verb (compare *his four trips to Brazil*). We should probably also list as this type of choice the insertion of the copula verb BE, which syntactically holds adjectives in predicative position, even though the term *choice* does not apply fully here (since English grammar makes use of only one copula).

In a **$_{\rightarrow}$LF(DL) choice**, a lexical function is introduced in order to express a special type of meaning, that we term *COLLOCATIONAL MEANING*. It is a meaning which tends to be verbalized in natural languages by means of collocations. A very clear example is intensification, which is expressed by the lexical function **Magn** (from Latin *magnus* 'big'): compare *deep*/*profound sorrow* and *high*/*great expectation*. On identifying a collocational meaning such as intensification in the semantic representation, the natural lexicalization model has to express it in the deep-syntactic tree by means of the corresponding lexical function. It is only when the choice of the base of the collocation (i.e. the argument of the lexical function) is finalized that a value for the collocate will be chosen (see next transition). The use of lexical functions in lexicalization provides us with a powerful means for modeling paraphrastic relations between sentences, thus allowing us to account for the multiplicity of choices a speaker is faced with in his use of the grammar and lexicon of any natural language. It also has practical advantages for applications in text generation systems, as demonstrated in Iordanskaja *et al.* (1996) —more on this in the Conclusion.

In a **$_{\rightarrow}$LF(DL)$_{\leftarrow}$ choice**, a lexical function is chosen both in order to express a collocational meaning and to hold part of the deep-syntactic structure of the sentence. Such is the case of the **Real$_1$** lexical function: it expresses the collocational meanings ‹to realize/to act accordingly/to make use/…› (already alluded to in Section 3.1) while providing us with a verbal governor for the sentence. For instance, **Real$_1$**(*requirement*) is *to fulfill*/*meet*/*satisfy* (*a requirement*), **Real$_1$**(*duty*) is *to do*/*carry out* (*one's duty*), etc.

In a **LF(DL)$_{\leftarrow}$ choice**, a lexical function is chosen strictly for syntactic reasons. Such is the case in the choice of **Oper$_1$**(TRIP) in our example (i.e. the choice of the following set of verbal collocates for *trip* {*to take* [ART ~], *to make* [ART ~], *to go* [*on* ART ~]}). It is not used in order to express any

meaning but in order to provide the syntactic tree with a verbal governor in the case where the meaning ‘go’ is expressed by TRIP.

4. THE DEEP SYNTACTIC-SURFACE SYNTACTIC TRANSITION

This last transition implements the choice of the actual lexical units to appear in the speaker's sentence, starting from already selected deep lexical units and LF applications. Only during deep syntactic-surface syntactic transition are lexical choices considered final.

4.1. Surface-syntactic representation: a tree hierarchy of all lexical units of the sentence

Two new types of lexical units can appear in the *SURFACE-SYNTACTIC REPRESENTATION* (which is the target level of this last transition): grammatical and collocational lexical units.

GRAMMATICAL LEXICAL UNITS belong to small subsets of the lexicon which are directly connected to the grammar of the language. In English, these are determiners (needed in many cases of noun phrase construction), pronouns, auxiliaries, governed prepositions (such as *on* in *to rely on someone*, as opposed to the deep lexical *on* in *the book on the table*), etc.

COLLOCATIONAL LEXICAL UNITS express collocational meanings (intensification, etc.) corresponding to lexical functions and can be viewed as some sort of "degenerate" deep lexical units. Such is the case of *to pay* in *to pay a visit*, or of *heavy* in *heavy bombardment*. The choice of a collocational lexical unit has to be performed at the latest stage as it is contingent upon the choice of the deep lexical unit that is the base of the collocation.

The grammatical lexical units we identify in (3b) are A, THE, OF and TO; there is only one collocational lexical unit, MAKE (the value selected for $\mathbf{Oper_1}$(TRIP)). The surface-syntactic representation of (3b) —see Figure 4 below— features all these units together with the deep lexical units already selected during the semantic transition.

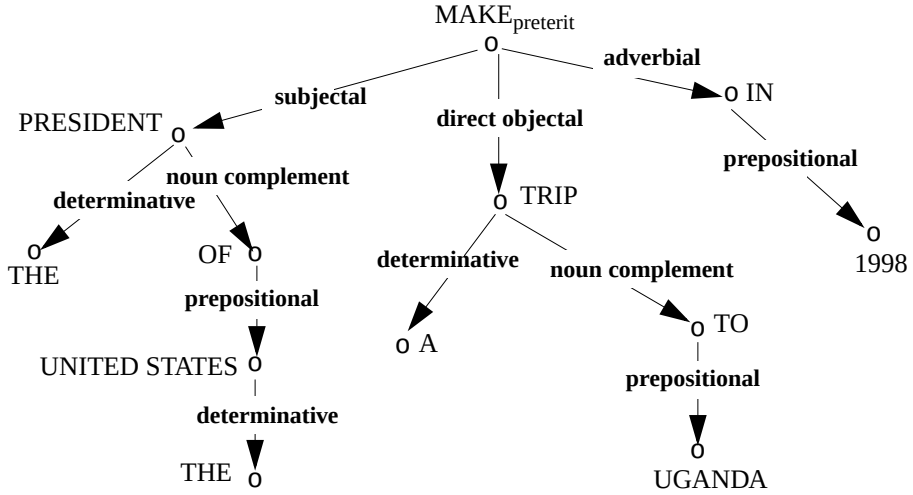

FIGURE 4. Surface-syntactic representation for (3b)

As in the deep-syntactic representation, we use here a dependency tree to encode the surface-syntactic structure of (3b). Notice that the dependency arcs are labelled with names of surface-syntactic relations (subjectal, determinative, etc.), which are language-specific.

We want to stress the fact that the use of particular grammatical formalisms, at all the levels of representation that we consider, does not imply that our lexicalization model can only be implemented based on these specific types of formal structures. We believe these formalisms to be partic-

ularly well-suited to support the kind of stratificational lexicalization we advocate, but they are by no means the only available option.

4.2. Choices performed during the deep syntactic-surface syntactic transition

As two types of lexical units can be introduced here —grammatical lexical units (GL) and collocational lexical units (CL)— we obtain six possible types of choices: $_{\rightarrow}$GL, $_{\rightarrow}$GL$_{\leftarrow}$, GL$_{\leftarrow}$, $_{\rightarrow}$CL, $_{\rightarrow}$CL$_{\leftarrow}$ and CL$_{\leftarrow}$.

In a **$_{\rightarrow}$GL choice**, a GL is introduced only in order to express a deep-syntactic content. Such could be the case for the subject pronoun in *It is on the table*, if we consider this sentence as being an alternative for *This object is on the table*. Besides "free" pronominalization (i.e. pronominalization which is not imposed by the grammar of the language), this type of choice does not seem to be very common.

In a **$_{\rightarrow}$GL$_{\leftarrow}$ choice**, a GL is selected both to express a deep-syntactic content and to play a given structural role. In English, auxiliary verbs are normally selected this way when they are part of the analytical form of a grammatical tense —for instance, *have* in *They have done it*. The choice of articles in (3b) falls also into this category: articles express, at the surface-syntactic level, the definite/indefinite grammatical feature associated with nominal deep lexical units in the deep-syntactic representation and their introduction has to be considered during synthesis in order to build grammatical English noun phrases. By this, we mean that even if the resulting noun phrase does not possess a determiner (for instance, in ***Work** is good for you*), the very absence of it is a means of expressing a specific grammatical meaning. In French, the situation is even clearer as noun phrases necessarily possess a determiner, unless they are part of some specific collocational expressions (e.g. Fr. *faire peur* lit. ‘to cause-fear’).

In a **GL$_{\leftarrow}$ choice**, a GL is selected for strictly surface-syntactic reasons. The standard example of such a choice in English is the selection of governed prepositions in prepositional phrases expressing verb or noun complements —see *of* and *to* in *president of the United States*, *trip to Uganda*, *go to Uganda*, etc. This type of choice is performed directly under the control of the governor's *SUBCATEGORIZATION FRAME* (also known as government pattern). In a lexicon used by a text generation system, the subcategorization frame is strictly speaking a lexical rule: a rule associated with a given lexical unit.

In a **$_{\rightarrow}$CL choice**, a CL is chosen solely in order to express an LF(DL) —i.e. to express the application of a lexical function to a deep lexical unit that is the base of a collocation. This can happen for instance when selecting a given **Magn**(*sorrow*) among all possible values (*deep*, *intense*, *profound*, etc.).

In a **$_{\rightarrow}$CL$_{\leftarrow}$ choice**, surface-syntactic considerations interfere in order to control the selection of a proper value for an LF(DL). For instance, among two possible values for **Oper$_1$**(*vacation*) —*to be* and *to spend*—only the former can be used in the absence of a locative complement: *He is on vacation in the States*/*He is spending his vacation in the States* vs. *He is on vacation*/**He is spending his vacation*.

In a **CL$_{\leftarrow}$ choice**, an LF(DL) is chosen in order to hold part of the surface-syntactic structure. It proves quite difficult to find illustrations of this type of choice. It could happen with verbs which possess a locative complement. The preposition introducing this complement cannot be specified in the subcategorization frame of the verb as its choice is contingent upon the choice of the noun the preposition will govern. Such is the case of the verb *to live* (*somewhere*); we say *He lives **in** an apartment* vs. *He lives **on** a farm*. The subcategorization frame for this verb has to be specified using the lexical function **Loc$_{\textbf{in}}$**, which returns locative prepositions contingent upon the noun expressing where the person lives: N_1 *to live* **Loc$_{\textbf{in}}$**(N_2)+N_2. Because it corresponds to a case of

standard DL, N_2 will be selected at the deep-syntactic level; the selection of $\mathbf{Loc_{in}}(N_2)$, however, is contingent upon the choice of N_2 and has to be performed strictly in order to build a correct grammatical structure for the second complement of the verb. The case of *to live* is by no means marginal and dozens of similar examples could be found; to illustrate this point, we will give one more illustration:

(7) *At 2:00PM, a team of experts arrived* ***in*** *the conference room.*
at *the ticket counter.*
on *Mars.*

It could be argued however that locative prepositions are not semantically empty, thus forcing us to consider these examples cases of $_{\rightarrow}\mathbf{CL}_{\leftarrow}$ **choices**. We could easily agree with this and we have not found until now undisputable cases of $\mathbf{CL}_{\leftarrow}$ **choices**.

5. CONCLUSION

The lexicalization model we have introduced presents the advantage of integrating in a unified model all the main types of lexicon-related choices that are involved in language generation. A crucial problem we had to disregard here is that of the selection of the initial message content. Where do meanings come from? How are they selected? We believe however that one of the main advantages of our lexicalization model is that it allows us to clearly distinguish between what concerns the lexicalization problem proper and what concerns the problem of building a message. This latter is too complex a task to be considered here. However, it should be emphasized that we are in favor of a minimal specification of the message that feeds into the lexicalization model. As mentioned throughout the paper with specific examples of calls to "extra-linguistic modules," we believe that natural languages force us, while performing lexical and grammatical choices, to express information that we would not have communicated otherwise. In order to model this, it is important to leave it to the surface generator to send requests for additional information in the process of forging a grammatical sentence (for similar proposals, see McKeown *et al.*1995, Nicolov *et al.* 1997 and Zock 1996).

Another point that should be discussed is whether the proposed model is indeed useful for designing actual text generation systems. We have already experimented in the past with some of the ideas presented here. The FoG system, for the generation of bilingual (English and French) weather forecasts (Kittredge and Polguère 1991; Kittredge *et al.* 1994), implements the selection of prepositions by means of subcategorization frames pretty much like the $GL_{\leftarrow}$ choice described above, as opposed to the selection of deep lexical units, which is performed at an earlier stage. However, this is not necessarily very original as it is now not uncommon to see text generation systems using more sophisticated lexicons, whose entries feature at least subactegorization frames of lexical units. Much more innovative was the experiment that was done with CL choices in the LFS system, for the bilingual (English and French) generation of statistical reports on the Canadian economy (Iordanskaja *et al.* 1996). This experiment, while showing how difficult the implementation of lexical functions can be from a procedural point of view, clearly demonstrated that a greater atomization of various choices involved in lexicalization is the key to the "clean" implementation of the generation of stylistically rich texts, where equivalent messages are not always expressed by identical linguistic means.

Apart from these previous projects, a full-fledged implementation of the proposed lexicalization model still remains to be done. It should be mentioned however that we are presently developing a prototype of a surface generator for French based on our lexicalization strategy, drawing from past implementations of Meaning-Text models for French and English.

ACKNOWLEDGMENTS

This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant n° OGPO200713. Many thanks to François Lareau, Helen Lim and reviewers for the SNLP'2000 conference for their very insightful comments. This text is a reworked and rewritten English version of Polguère (1998), on which we had received feedback from L. Danlos, S. Kahane and M. Zock.

REFERENCES

BATEMAN, J. A. 1998. Automated Discourse Generation. *In* Encyclopedia of Library and Information Science, vol. 62, supplement 25. *Edited by* A. Kent. Dekker, New York, pp. 1-54.

DE SMEDT, K. 1990. IPF: An Incremental Parallel Formulator. *In* Current Research in Natural Language Generation. *Edited by* R. Dale, C. S. Mellish, & M. Zock. Academic Press, London, pp. 167-192.

ELHADAD, M., K. MCKEOWN, and J. ROBIN. 1997. Floating Constraints in Lexical Choice. Computational Linguistics, **23**(2): 195-239.

IORDANSKAJA, L., M. KIM, and A. POLGUÈRE. 1996. Some Procedural Problems in the Implementation of Lexical Functions for Text Generation. *In* Wanner (1996), pp. 279-297.

IORDANSKAJA, L., R. I. KITTREDGE, and A. POLGUÈRE. 1991. Lexical selection and paraphrase in a Meaning-Text generation model. *In* Natural Language Generation in Artificial Intelligence and Computational Linguistics. *Edited by* C. L. Paris, W. R. Swartout, & W. C. Mann. Kluwer, Dordrecht, pp. 293-312.

KITTREDGE, R. I., E. GOLDBERG, M. KIM, and A. POLGUÈRE. 1994. Sublanguage Engineering in the FoG System (Poster paper). *In* Proceedings of the Fourth Conference on Applied Natural Language Processing, Stuttgart, Germany, pp. 215-216.

KITTREDGE, R. I., and A. POLGUÈRE. 1991. Dependency Grammars for Bilingual Text Generation: Inside FoG's Stratificational Models. *In* Proceedings of the International Conference on Current Issues in Computational Linguistics, Penang, Malaysia, pp. 318-330.

MCKEOWN, K. R. 1985. Text Generation: Using discourse strategies and focus constraints to generate natural language text. Cambridge University Press, Cambridge.

MCKEOWN, K. R., J. ROBIN, and K. KUKICH. 1995. Generating Concise Natural Language Summaries. Information Processing & Management, **31**(5): 703-733.

MEL'CUK, I. A. 1981. Meaning-Text Models: A Recent Trend in Soviet Linguistics. Annual Review of Anthropology, **10**: 27-62.

MEL'CUK, I. A. 1988. Dependency Syntax: Theory and Practice. State University of New York Press, Albany.

MEL'CUK, I. A. 1995. Phrasemes in Language and Phraseology in Linguistics. *In* Idioms: Structural and Psychological Perspectives. *Edited by* M. Everaert, E.-J. van der Linden, A. Schenk, & Rob Schrender. Erlbaum, Hillsdale/Hove, pp. 167-232.

MEL'CUK, I. A., and A. POLGUÈRE. 1987. A Formal Lexicon in the Meaning-Text Theory (or How to Do Lexica with Words). Computational Linguistics, **13**(3-4): 261-275.

NICOLOV, N., C. MELLISH, and G. RICHIE. 1997. Approximate Chart Generation from Non-Hierarchical Representations. *In* Recent Advances in Natural Language Processing: Selected Papers from RANLP'95. *Edited by* R. Mitkov, & N. Nicolov. Benjamins, Amsterdam, pp. 273-294.

POLGUÈRE, A. 1998. Pour un modèle stratifié de la lexicalisation en génération de texte. Traitement Automatique des Langues (T.A.L.), **39**(2): 57-76.

POLGUÈRE, A. 2000. Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. *To appear in* Proceedings of EURALEX'2000, Stuttgart, Germany.

STEDE, M. 1995. Lexicalization in natural language generation: a survey. Artificial Intelligence Review, **8**: 309-336.

STEDE, M. 1996. Lexical semantics and knowledge representation in multilingual sentence generation. PhD dissertation, University of Toronto, Toronto, Canada.

STEDE, M. 1998. A Generative Perspective on Verb Alternations. Computational Linguistics (Special Issue on Natural Language Generation), **24**(3): 401-430.

SOWA, J. F. 1984. Conceptual structures: information processing in mind and machine. Addison-Wesley, Reading/Don Mills.

TANAKA-ISHII, K., K. HASIDA, and I. NODA. 1998. Reactive Content Selection in the Generation of Real-time Soccer Commentary. *In* Proceedings of the 17th International Conference on Computational Linguistics (COLING'98), vol. 2. Montreal, Canada, pp. 1282-1288.

TESNIÈRE, L. 1959. Éléments de syntaxe structurale. Klincksieck, Paris.

WANNER, L.(ed.) 1996. Lexical Functions in Lexicography and Natural Language Processing. Benjamins, Amsterdam.

WANNER, L. 1997. Exploring Lexical Resources for Text Generation in a Systemic Functional Language Model. PhD dissertation. Universität des Saarlandes, Saarbrücken, Germany.

WANNER, L. 1999. On the representation of collocations in a multilingual computational lexicon. Traitement Automatique des Langues (T.A.L.), **40**(1): 55-86.

ZOCK, M. 1996. The Power of Words in Message Planning. *In* Proceedings of the 16th International Conference on Computational Linguistics (COLING'96). Copenhagen, Denmark, pp. 990-995.