Lexical Selection in the Process of Language Generation

James Pustejovsky
Department of Computer Science
Brandeis University
Waltham, MA 02254
617-736-2709
jamesp@brandeis.csnet-relay

Sergei Nirenburg
Computer Science Department
Carnegie-Mellon University
Pittsburgh, PA. 15213
412-268-3823
sergei@cad.cs.cmu.edu

Abstract

In this paper we argue that lexical selection plays a more important role in the generation process than has commonly been assumed. To stress the importance of lexical-semantic input to generation, we explore the distinction and treatment of generating *open* and *closed class* lexical items, and suggest an additional classification of the latter into *discourse-oriented* and *proposition-oriented* items. Finally, we discuss how lexical selection is influenced by thematic (*focus*) information in the input.

1. Introduction

There is a consensus among computational linguists that a comprehensive analyzer for natural language must have the capability for robust lexical **disambiguation**, i.e., its central task is to select appropriate meanings of lexical items in the input and come up with a non contradictory, unambiguous representation of both the propositional and the non-propositional meaning of the input text. The task of a natural language generator is, in some sense, the opposite task of rendering an unambiguous meaning in a natural language. The main task here is to to perform principled **selection** of a) lexical items and b) the syntactic structure for input constituents, based on lexical semantic, pragmatic and discourse clues available in the input. In this paper we will discuss the problem of **lexical** selection.

The problem of selecting lexical items in the process of natural language generation has not received as much attention as the problems associated with expressing explicit grammatical knowledge and control. In most of the generation systems, lexical selection could not be a primary concern due to the overwhelming complexity of the generation problem itself. Thus, MUMBLE concentrates on grammar-intensive control decisions (McDonald and Pustejovsky, 1985a) and some stylistic considerations (McDonald and Pustejovsky, 1985b); TEXT (McKeown, 1985) stresses the strategical level of control decisions about the overall textual shape of the generation output. [1]KAMP (Appelt, 1985) emphasizes the role that dynamic planning plays in controlling the process of generation, and specifically, of referring expressions; NIGEL (Mann and Matthiessen, 1983) derives its control structures from the choice systems of systemic grammar, concentrating on grammatical knowledge without fully realizing the 'delicate' choices between elements of what systemicists call *lexis* (e.g., Halliday, 1961). Thus, the survey in Cumming (1986) deals predominantly with the grammatical aspects of the lexicon.

We discuss here the problem of lexical selection and explore the types of control knowledge that are necessary for it. In particular, we propose different control strategies and epistemological foundations for the selection of members of a) open-class and b) closed-class lexical items. One of the most important aspects of control knowledge our generator employs for lexical selection is the non-propositional information (including knowledge about focus and discourse cohesion markers). Our generation system incorporates the discourse and textual knowledge provided by TEXT as well as the power of MUMBLE's grammatical constraints and adds principled lexical selection (based on a large semantic knowledge base) and a control structure capitalizing on the inherent flexibility of distributed architectures. [2] The specific innovations discussed in this paper are:

[1] Derr and McKeown, 1984 and McKeown, 1985, however, discuss thematic information, i.e. focus, as a basis for the selection of anaphoric pronouns. This is a fruitful direction, and we attempt to extend it for treatment of additional discourse-based phenomena.

[2] Rubinoff (1986) is one attempt at integrating the textual component of TEXT with the grammar of MUMBLE. This interesting idea leads to a significant improvement in the performance of sentence production. Our approach differs from this effort in two important repsects. First, in Rubinoff's system the output of TEXT serves as the input to MUMBLE, resulting in a cascaded process. We propose a distributed control where the separate knowledge sources contribute to the control when they can, opportunistically. Secondly, we view the generation process as the product of many more components than the number proposed in current generators. For a detailed discussion of these see Nirenburg and Pustejovsky, in preparation.

1. We attach importance to the question of what the input to a generator should be, both as regards its content and its form; thus, we maintain that discourse and pragmatic information is absolutely essential in order for the generator to be able to handle a large class of lexical phenomena; we distinguish two sources of knowledge for lexical selection, one discourse and pragmatics-based, the other lexical semantic.

2. We argue that lexical selection is not just a side effect of grammatical decisions but rather acts to flexibly constrain concurrent and later generation decisions of either lexical or grammatical type.

For comparison, MUMBLE lexical selections are performed after some grammatical constraints have been used to determine the surface syntactic structure; this type of control of the generation process does not seem optimal or sufficient for all generation tasks, although it may be appropriate for on-line generation models; ; we argue that the decision process is greatly enhanced by making lexical choices early on in the process. Note that the above does not presuppose that the control structure for generation is to be like cascaded transducers; in fact, the actual system that we are building based on these principles, features a distributed architecture that supports non-rigid decision making (it follows that the lexical and grammatical decisions are not explicitly ordered with respect to each other). This architecture is discussed in detail in Nirenburg and Pustejovsky, in preparation.

3. We introduce an important distinction between open-class and closed-class lexical items in the way they are *represented* as well as the way they are *processed* by our generator; our computational, processing-oriented paradigm has led us to develop a finer classification of the closed-class items than that traditionally acknowledged in the psycholinguistic literature; thus, we distinguish between discourse oriented closed-class (DOCC) items and proposition oriented ones (POCC);

4. We upgrade the importance of knowledge about focus in the sentence to be generated so that it becomes one of the prime heuristics for controlling the entire generation process, including both lexical selection and grammatical phrasing.

5. We suggest a comprehensive design for the concept lexicon component used by the generator, which is perceived as a combination of a general-purpose semantic knowledge base describing a subject domain (a subworld) and a generation-specific lexicon (indexed by concepts in this knowledge base) that consists of a large set of discrimination nets with semantic and pragmatic tests on their nodes.

These discrimination nets are distinct from the choosers in NIGEL's choice systems, where grammatical knowledge is not systematically separated from the lexical semantic knowledge (for a discussion of problems inherent in this approach see McDonald, Vaughan and Pustejovsky, 1986); the pragmatic nature of some of the tests, as well as the fine level of detail of knowledge representation is what distinguishes our approach from previous conceptual generators, notably PHRED (Jacobs, 1985)).

2. Input to Generation

As in McKeown (1985,1986) the input to the process of generation includes information about the discourse within which the proposition is to be generated. In our system the following static knowledge sources constitute the input to generation:

1. A representation of the meaning of the text to be generated, chunked into proposition-size modules, each of which carries its own set of contextual values; (cf. TRANSLATOR, Nirenburg et al., 1986, 1987);
2. the semantic knowledge base (concept lexicon) that contains information about the types of concepts (objects (mental, physical and perceptual) and processes (states and actions)) in the subject domain, represented with the help of the description module (DRL) of the TRANSLATOR knowledge representation language. The organizational basis for the semantic knowledge base is an empirically derived set of inheritance networks (isa, made-of, belongs-to, has-as-part, etc.).
3. The specific lexicon for generation, which takes the form of a set of discrimination nets, whose leaves are marked with lexical units or lexical gaps and whose non-leaf nodes contain discrimination criteria that for open-class items are derived from selectional restrictions, in the sense of Katz and Fodor (1963) or Chomsky (1965), as modified by the ideas of preference semantics (Wilks, 1975, 1978). Note that most closed-class items have a special status in this generation lexicon: the discrimination nets for them are indexed not by concepts in the concept lexicon, but rather by the types of values in certain (mostly, nonpropositional) slots in input frames;
4. The history of processing, structured along the lines of the *episodic memory organization* suggested by Kolodner (1984) and including the feedback of the results of actual lexical choices during the generation of previous sentences in a text.

3. Lexical Classes

The distinction between the open- and closed-class lexical units has proved an important one in psychology and psycholinguistics. The manner in which retrieval of elements from these two classes operates is taken as evidence for a particular mental lexicon structure. A recent proposal (Morrow, 1986) goes even further to explain some of our discourse processing capabilities in terms of the properties of some closed-class lexical items. It is interesting that for this end Morrow assumes, quite uncritically, the standard division between closed- and open-class lexical categories: 'Open-class categories include content words, such as nouns, verbs and adjectives... Closed-class categories include function words, such as articles and prepositions...' (op. cit., p. 423). We do not elaborate on the definition of the open-class lexical items. We have, however, found it useful to actually define a particular subset of closed-class items as being discourse-oriented, distinct from those closed-class items whose processing does not depend on discourse knowledge.

A more complete list of closed-class lexical items will include the following:

- determiners and demonstratives (*a, the, this, that*);
- quantifiers (*most, every, each, all of*);
- pronouns (*he, her, its*);
- deictic terms and indexicals (*here, now, I, there*);
- prepositions (*on, during, against*);
- parentheticals and attitudinals (*as a matter of fact, on the contrary*);
- conjunctions, including discontinuous ones (*and, because, neither...nor*);
- primary verbs (*do, have, be*);
- modal verbs (*shall, might, ought to*);
- *wh*-words (*who, why, how*);
- expletives (*no, yes, maybe*).

We have concluded that the above is not a homogeneous list; its members can be characterized on the basis of what knowledge sources are used to evaluate them in the generation process. We have established two such distinct knowledge sources: purely propositional information and contextual and discourse knowledge. Those closed-class items that are assigned a denotation only in the context of an utterance will be termed discourse-oriented closed class (DOCC) items; this includes determiners, pronouns, indexicals, and temporal prepositions. Those contributing to the propositional content of the utterance will be called proposition-oriented closed-class (POCC) items. These include modals, locative and function prepositions, and primary verbs.

According to this classification, the "definiteness effect" (that is, whether a definite or an intefinite noun phrase is selected for generation) is distinct from general quantification, which appears to be decided on the basis of propositional factors. Note that prepositions no longer form a natural class of simple closed-class items. For example, in (1) the preposition *before* unites two entities connected through a discourse marker. In (2) the choice of the preposition *on* is determined by information contained in the propositional content of the sentence.

(1) John ate breakfast *before* leaving for work.

(2) John sat *on* the bed.

We will now suggest a set of processing heuristics for the lexical selection of a member from each lexical class. This classification entails that the lexicon for generation will contain only open-class lexical items, because the rest of the lexical items do not have an independent epistemological status, outside the context of an utterance. The selection of closed-class items, therefore, comes as a result of the use of the various control heuristics that guide the process of generation. In other words, they are incorporated in the procedural knowledge rather than the static knowledge.

4.0 Lexical Selection

4.1 Selection of Open-Class Items

A significant problem in lexical selection of open-class items is how well the concept to be generated matches the desired lexical output. In other words, the input to generate in English the concept 'son's wife's mother' will find no single lexical item covering the entire expression. In Russian, however, this meaning is covered by a single word 'svatja.' This illustrates the general problem of **lexical gaps** and bears on the question of how strongly the conceptual representation is influenced by the native tongue of the knowledge-engineer. The representation must be comprehensive yet flexible enough to accommodate this kind of problem. The processor, on the other hand, must be constructed so that it can accommodate lexical gaps by being able to build the most appropriate phrase to insert in the slot for which no single lexical unit can be selected (perhaps, along the lines of McDonald and Pustejovsky, 1985a).

To illustrate the knowledge that bears upon the choice of an open-class lexical item, let us trace the process of lexical selection of one of the words from the list: *desk, table, dining table, coffee table, utility table.* Suppose, during a run of our generator we have already generated the following partial sentence:

(3) *John bought a .....*

and the pending input is as partially shown in Figures 1-3. Figure 1 contains the instance of a concept to be generated.

```
(stol34
   (instance-of stol)
   (color black)
   (size small)
   (height average)
   (mass average)
   (made-of steel)
   (location-of eat))
```

Figure 1

```
(stol
   (isa furniture)
   (color black brown yellow white)
   (size small average)
   (height low average high)
   (mass less-than-average average)
   (made-of wood plastic steel)
   (location-of eat write sew work)
   (has-as-part ( leg leg leg (leg) top)
       (topology ON (top leg)))
```

Figure 2

Figure 2 contains the representation of the corresponding type in the semantic knowledge base. Figure 3 contains an excerpt from the English generation lexicon, which is the discrimination net for the concept in Figure 2.

```
case location-of of
      eat:  case height of
                  low:    coffee table
                  average:  dining table
      write:   desk
      sew:   sewing table
      saw:   workbench
      otherwise:   table
```

Figure 3

In order to select the appropriate lexicalization the generator has to traverse the discrimination net, having first found the answers to tests on its nodes in the representation of the concept token (in Figure 1). In addition, the latter representation is compared with the representation of the concept type and if non-default values are found in some slots, then the result of the generation will be a noun phrase with the above noun as its head and a number of adjectival modifiers. Thus, in our example, the generator will produce 'black steel dining table'.

4.2 Selection of POCC Items

Now let us discuss the process of generating a proposition oriented lexical item. The example we will use here is that of the function preposition *to*. The observation here is that if *to* is a POCC item, the information required for generating it should be contained within the propositional content of the input representation; no contextual information should be necessary for the lexical decision. Assume that we wish to generate sentence (1) where we are focussing on the selection of *to*.

(1) John walked *to* the store.

If the input to the generator is

```
(walk
  (Actor John)
  (Location "here")
  (Source U)
  (Goal store23)
  (Time past2)
  (intention U)
  (Direction store23))
```

then the only information necessary to generate the preposition is the case role for the goal, *store*. Notice that a change in the lexicalization of this attribute would only arise with a different input to the generator. Thus, if the goal were unspecified, we might generate (2) instead of (1); but here the propositional content is different.

(2) John walked towards the store.

In the complete paper we will discuss the generation of two other DOCC items; namely, quantifiers and primary verbs, such as *do* and *have*.

4.2 Selection of DOCC Items: Generating a discourse anaphor

Suppose we wish to generate an anaphoric pronoun for an NP in a discourse where its antecedent was mentioned in a previous sentence. We illustrate this in Figure 2. Unlike open-class items, pronominals are not going to be directly associated with concepts in the semantic knwoledge base. Rather, they are generated as a result of decisions involving contextual knowledge, the beliefs of the speaker and hearer, and previous utterances. Suppose, we have already generated (4) and the next sentence to be generated also refers to the same individual and informs us that John was at his father's for two days.

(1) John$_i$ visited his father.

(2) *He*$_i$ stayed for two days.

Immediate focus information, in the sense of Grosz (1979) interacts with a history of the previous sentence structures to determine a strategy for selecting the appropriate anaphor. Thus, selecting the appropriate pronoun is an attached procedure. The heuristic for discourse-directed pronominalization is as follows:

IF: (1) the input for the generation of a sentence includes an instance of an object present in a recent input; and
(2) the the previous instance of this object (the potential antecedent) is in the topic position; and
(3) there are few intervening potential antecedents; and
(4) there is no focus shift in the space between the occurrence of the antecedent and the current object instance

THEN: realize the current instance of that object as a pronoun; consult the grammatical knowledge source for the proper gender, number and case form of the pronoun.

In McDonald and Pustejovsky (1985b) a heursitic was given for deciding when to generate a full NP and when a pronoun. This decision was fully integrated into the grammatical decisions made by MUMBLE in terms of realization-classes, and was no different from the decision to make a sentence active or passive. Here, we are separating discourse information from linguistic knowledge. Our system is closer to McKeown's (1985, 1986) TEXT system, where discourse information acts to constrain the control regimen for linguistic generation. We extend McKeown's idea, however, in that we view the process of lexical selection as a constraining factor *in general.* In the complete paper, we illustrate how this works with other discourse oriented closed-class items.

5. The Role of Focus in Lexical Selection

As witnessed in the previous section, focus is an important factor in the generation of discourse anaphors. In this section we demonstrate that focus plays an important role in selecting non-discourse items as well. Suppose your generator has to describe a financial transaction as a result of which

(1) Bill is the owner of a car that previously belonged to John, and
(2) John is richer by $2,000.

Assuming your generator is capable of representing the grammatical structure of the resulting English sentence, it still faces an important decision of how to express lexically the actual transaction relation. Its choice is to either use *buy* or *sell* as the main predicate, leading to either (1) or (2), or to use a non-perspective phrasing where neither verb is used.

(1) Bill bought a car from John for $2,000.
(2) John sold a car to Bill for $2,000.

We distinguish the following major contributing factors for selecting one verb over the other; (1) the intended perspective of the situation, (2) the emphasis of one activity rather than another, (3) the focus being on a particular individual, and (4) previous lexicalizations of the concept.

These observations are captured by allowing *focus* to operate over several expression including event-types such as *transfer.* Thus, the variables at play for focus include:

- end-of-transfer,
- beginning-of-transfer,
- activity-of-transfer,
- goal-of-object,
- source-of-object,
- goal-of-money,
- source-of-money.

That is, lexicalization depends on which expressions are in focus. For example, if John is the immediate focus (as in McKeown (1985)) and beginning-of-transfer is the current-focus, the generator will lexicalize from the perspective of the selling, namely (2). Given a different focus configuration in the input to the generator, the selection would be different and another verb would be generated.

6. Conclusion

In this paper we have argued that lexical selection is an important contributing factor to the process of generation, and not just a side effect of grammatical decisions. Furthermore, we claim that open-class items are not only conceptually different from closed-class items, but are processed differently as well. Closed class items have no epistemological status other than procedural attachments to conceptual and discourse information. Related to this, we discovered an interesting distinction between two types of closed-class items, distinguished by the knowledge sources necessary to generate them; discourse oriented and proposition-oriented. Finally, we extend the importance of focus information for directing the generation process.

References

[1] Appelt, Douglas *Planning English Sentences*, Cambridge U. Press.

[2] Chomsky, Noam *Aspects on the Theory of Syntax*, MIT Press.

[3] Cumming, Susanna, "A Guide to Lexical Acquisition in the JANUS System" ISI Research Report ISI/RR-85-162, Information Sciences Institute, Marina del Rey, California, 1986a.

[4] Cumming, Sussana, "The Distribution of Lexical Information in Text Generation", presented for Workshop on Automating the Lexicon, Pisa, 1986b.

[5] Derr, K. and K. McKeown "Focus in Generation, COLING 1984

[6] Dowty, David R., Word Meaning and Montague Grammar, D. Reidel, Dordrecht, Holland, 1979.

[7] Halliday, M.A.K. "Options and functions in the English clause". *Brno Studies in English* 8, 82-88.

[8] Jacobs, Paul S., "PHRED: A Generator for Natural Language Interface", Computational Linguistics, Volume 11, Number 4, 1985.

[9] Katz, Jerrold and Jerry A. Fodor, "The Structure of a Semantic Theory", Language Vol 39, pp.170-210, 1963.

[10] Mann, William and Matthiessen, "NIGEL: a Systemic Grammar for Text Generation", in Freddle (ed.), *Systemic Perspectives on Discourse*, Ablex.

[11] McDonald, David and James Pustejovsky, "Description directed Natural Language Generation" Proceedings of IJCAI-85. Kaufmann.

[12] McDonald, David and James Pustejovsky, "A Computational Theory of Prose Style for Natural Language Generation, Proceedings of the European ACL, University of Geneva, 1985.

[13] McKeown, Kathy *Text Generation*, Cambridge University Press.

[14] McKeown, Kathy, "Stratagies and Constraints for Generating Natural Language Text", in Bolc and McDonald, 1987.

[15] Morrow "The Processing of Closed Class Lexical Items", in Cognitive Science 10.4, 1986.

[16] Nirenburg, Sergei, Victor Raskin, and Allen Tucker, "The Structure of Interlingua in TRANSLATOR", in Nirenburg (ed.) *Machine Translation: Theoretical and Methodological Issues*, Cambridge University Press. 1987.

[17] Wilks, Yorick "Preference Semantics," *Artificial Intelligence*, 1975.