Lexical choice in context: generating procedural texts

Agnès Tutin[1], Richard Kittredge
Département de linguistique,
Université de Montréal
C.P. 6128, Succ "A",
Montréal P.Q. H3C 3J7 Canada

Abstract

This paper shows how lexical choice during text generation depends on linguistic context. We argue that making correct lexical choice in the textual context requires distinguishing properties of concepts, which are more or less independent of the language, from language-specific representations of text where lexemes and their semantic and syntactic relations are represented. In particular, Lexical Functions are well-suited to formalizing anaphoric lexical links in text, including the introduction of superordinates. This sheds new light on the notion of "basic level", which has recently been applied to lexical selection in generation. Some constraints governing the generation of lexical and grammatical anaphora are proposed for procedural text, using examples from the sublanguage of recipes.

0. Introduction

Lexical choice cannot be made during text generation without taking into account the **linguistic context**, both the lexical context of immediately surrounding words and the larger textual context.

a) **Lexical context** consists of the words (or rather, the lexical specifications of nascent words) that enter into syntactic relations with the lexical item being generated. This intra-clausal context is crucial for formulating collocational constraints, which restrict ways of expressing a precise meaning to certain lexical items, for example as in expressions like *pay attention*, *receive attention* or *narrow escape*. The importance of collocational constraints has been emphasized in the literature on text generation and machine translation (Bateman & Wanner 1990, Iordanskaja *et al.* 1991, Nirenburg & Nirenburg 1988, Heid & Raab 1989).

b) **Textual context** consists of the linguistic content of previous and subsequent clauses. This context is the scope for cohesive links (Halliday & Hasan 1976) with the lexical items to be generated in the current clause.

The great majority of cohesive links are anaphoric in nature[2]. A textual element T is an anaphor with respect to an antecedent A (previously introduced in the text) if the semantic or referential interpretation of T[3] depends on the interpretation of A. When generating anaphors, it is therefore the previous context that must be taken into account, as in:

(1) Prepare the carrots, the celery and the asparagus.
Cook the vegetables in boiling water for ten minutes.

Two textual element are coreferential if they refer to the same extralinguistic reality. Coreferential elements in our examples are often written in italics, or indicated by identical subscripts.

Failure to choose an appropriate anaphoric expression during generation typically leads to awkward or unacceptable text such as (2):

(2) a. Prepare the carrots, the celery and the asparagus.
b. Cook the carrots, the celery and the asparagus in boiling water.
c. Take the carrots, the celery and the asparagus out after ten minutes.

In this paper, we examine the mechanisms required for making natural lexical choice as a function of preceding text and its reference to extralinguistic objects or concepts. In particular, we are interested in lexical anaphora, where open-class lexical items or expressions provide a coreference link to one or more such items in preceding clauses. For example, in (1) *vegetables* is a lexical coreferential anaphor of *the carrots, the celery and the asparagus*..

In what follows, we aim to show that correct lexicalization in context requires access to both the conceptual reference and the linguistic properties of preceding text. For a pipelined generation architecture which maps from abstract representation levels towards text, this implies distinguishing a conceptual level, more or less independent of the language, from language-specific representation levels which encode lexemes and the grammatical relation between them. In particular, we illustrate the paradigmatic Lexical Functions (hereafter LFs) of Mel'čuk's *Explanatory Combinatorial Dictionary* (hereafter ECD) (Mel'čuk et al. 1988; Mel'čuk & Polguère 1987).

1. Varieties of lexical anaphora

Before reviewing constraints on the introduction of lexical anaphora during generation, we give examples of important types of coreferential anaphoric links[4].

We consider an anaphor to be lexical only if we can establish a semantic link between the anaphor and its antecedent. Therefore, in the following example:

(3) *Edith Cresson* arrived Monday at 9:00. At 11:00, *the Prime Minister of France* gave a press conference.

---

[1]) Now in Centre National d'Études des Télécommunications, LAA/SLC/AIA, Route de Trégastel BP 40 22301 Lannion Cédex, France.

[2]) In the case of cataphora, as in the following sentence,

If you want to, peel and chop the potatoes.

the subsequent context must be taken into account.

[3]) Reference of a textual element is the association between a textual element and extra-linguistic reality.

[4]) We will not treat here non -coreferential anaphora like:
Marie threw away all her old dresses because she wanted to buy new ones.

Prime Minister of France is a cognitive coreferential anaphor of *Edith Cresson*, but not a lexical one because the coreferential link between the two phrases is based on world knowledge, and not on linguistic semantics.

One type of coreferential lexical anaphora is called "reiteration" by Halliday & Hasan (1976) with three subtypes: exact repetition, illustrated by (4b), synonym substitution (4b'), and superordinate substitution (4b"). We can add to this group partial repetition (4b"'):

(4) a. 1 bottle of light red bordeaux.
b. Pour the light red bordeaux on the meat.
b'. Pour the light red bordeaux wine on the meat.
b". Pour the light wine on the meat.
b"'. Pour the red bordeaux on the meat.

Nominalization provides another way of introducing a coreferential link to a previous verb:

(5) Cook the rabbit for two hours.
Ten minutes before the end of cooking, add the spices.

Coreferential lexical links can also be established between an action and its result.

(6) Meanwhile, mix the egg yolk$_i$ with the sugar$_j$.
Pour the milk on the mixture$_{i+j}$.

In this example, *mixture* has no direct semantic link with its antecedents *egg yolk* and *sugar*. The link appears indirectly through the verb *mix*.

Another type of lexical anaphora occurs with nouns denoting typical actants of an antecedent verb:

(7) Marga$_i$ was lecturing to third year students$_k$. The lecturer$_i$ was very interesting and the audience$_k$ quite attentive.

In this case, *lecturer* is linked coreferentially with *Marga* because it is the "agent noun" of *lecture*, while *audience* is the corresponding "patient noun", and is coreferential with *third year students*.

These examples illustrate some of the diversity of lexico-semantic resources needed to build coreferential links in text. Text generation therefore requires a lexicon which gives access to the full range of such resources from the "viewpoint" of the antecedent lexeme. As seen in the next section, LFs provide an appropriate access mechanism for choosing the correct anaphor.

2. Lexical Functions of the ECD for creating lexical anaphora

Lexical Functions of the ECD provide a formalism representing many common instances of coreferential anaphora. Formally defined, a Lexical Function f is a correspondence between a lexical item L, called the key word of f, and a set of lexical items f(L) - the values of f (Mel'čuk & Zholkovskij 1970, Mel'čuk 1988b). Approximately sixty standard Lexical Functions have been defined (for a recent description of LFs in the ECD in English, see Mel'čuk & Polguère 1987). They can be divided into two subsets: **syntagmatic LFs** and **paradigmatic LFs**.

- **Syntagmatic or collocational LFs** are used to link unpredictable lexical cooccurrences in texts between the key word and its values through a specific semantic relation. Typical examples of syntagmatic LFs are Oper$_i$ (semantically empty verb which takes the i-th actant[5] of the key word as its subject and the key word as its direct object), like Oper$_1$(*attention*) = *pay*, Oper$_2$(*attention*) = *receive* or Magn(*escape*) = *narrow*. These examples show that these LFs convey cooccurrence relations.

- **Paradigmatic LFs** are used to express semantic relations in the lexicon between the key word of the LF and its values, but not cooccurrence relations. Typical examples are S_1(*lecture*) = *lecturer* (S_1:Noun of the first typical actant), S_{loc} (*box*) = *ring* (S_{loc}: Noun of typical place), S_0(*buy*) = *purchase* (S_0: Derived noun). Some paradigmatic LFs can be used to analyse or generate lexical coreference relations:

- Syn: synonym	Syn(calling) = vocation
- $Conv_{ijkl}$: conversive	$Conv_{3214}$(sell) = buy
- Gener: generic word	Gener(apple) = fruit
- S_i: typical noun for the i-th actant	S_1(lecture) = lecturer
- S_{instr}: noun for typical instrument	Sinstr(paint) = brush
- S_{med}: noun for typical means	S_{med}([to]salt) = salt
- S_{loc}: noun for typical place	S_{loc}(box) = ring
- S_{res}: noun for typical result	S_{res}(mix) = mixture
- S_{mod}: noun for typical mode	Smod(write) = writing
- S_θ : name of action	S_0(buy) = purchase

Relations encoded by these LFs can appear in **direct coreferential relations** in texts when the value of the function and the key word maintain a semantic relationship directly formalizable through a LF such as S_{res}, Gener, Syn and $Conv_{ijkl}$, as in:

(8) Gener(*lamb*) = *meat*
Buy *lamb*. Be sure the *meat* is very fresh.

LFs can be used to formalize **indirect lexical coreference** when coreference exists between lexical items and a dependent. The dependent may be an actant as in (7) (*lecturer*, the S_1(*lecture*) is coreferential with the first actant *Marga* of *lecture* whereas *audience*, the S_2(*lecture*) is coreferential with the second actant of *lecture*), or an adverbial, as in the following example:

(9) S_{loc}(patiner) = patinoire
Marguerite et Jean ont patiné sur le canal$_i$ Rideau.
Cette patinoire$_i$ fait 8km de long.
[Marguerite and Jean skated on the Rideau Canal. This "skating rink" is 8 km long.]

In (9), *patinoire* , S_{loc}(*patiner*) is coreferential with *canal Rideau*.

Moreover, LFs can be combined, as we see in the following table:

[5])In the ECD, *lecture* will be described as a noun which has three syntactic actants: X's (actant I) *lecture* to Y (actant II) on Z (actant III), for example Jean's (actant I) *lecture* to third year students (actant II) on semantic causality (actant III).

LFs or composition of LFs	key word	values
Gener	achat [purchase]	transaction [deal]
Gener	vente [sale]	transaction [deal]
Gener	transaction [deal]	action [action]
Gener	auto [car]	véhicule [vehicle]
Gener	voiture [car]	véhicule [vehicle]
Syn	voiture [car]	auto [car]
$Conv_{3214}$	acheter [buy]	vendre [sell]
S_0	acheter [buy]	achat [purchase]
S_0	vendre [sell]	vente [sale]
Gener o Gener	achat [purchase]	action [action]
Gener o Syn	auto [car]	véhicule [vehicle]
Gener o $Conv_{3214}$	vente [sale]	transaction [deal]
Gener o S_0	acheter [buy]	transaction [deal]
S0 o $Conv_{3124}$	acheter [buy]	vente [sale]
$Conv_{3214}$ o S_0	acheter [buy]	vente [sale]

Table 1: LFs and compositions of LFs for direct coreference links

The following facts should be noted about compositions:
- Composition is not commutative. Thus, $S_0(Conv_{3214}(acheter)) = Conv_{3214}(S_0(acheter)) \approx$ *vente* but $Gener(S_0(acheter)) \neq S_0(Gener(acheter))$ because Gener(*acheter*) does not have a value.
- Some compositions are reducible. For example, the LF Syn plays a transparent role in composition.

In the perspective of text generation, this formalism appears very interesting for building coreferential expressions. To point back to a referent already introduced, LFs and compositions of LFs offer many possible ways for lexicalizing a given referent. For example, let us suppose that after having introduced the following sentence,

(10) a. Laisser étuver la viande. [Let the meat steam.]

we have to refer again to the action *la viande étuve*. We could try to use a nominalization (S_0). But, as there is no nominalization for the verb *étuver*, we could use instead the nominalization of the generic term, $S_0(Gener(étuver)) = cuisson$. We could thus produce the following sentence:

(11) b. A la fin de la cuisson, ajouter les épices
[At the end of cooking, add the spices]

In the next section, we will examine the case of a complex lexical anaphor: the superordinate term.

3. Superordinates and basic nouns

The use of superordinate terms as anaphors raises several interesting questions.

First, to the extent that a generic concept (for two or more specific concepts) has a simple expression in a language, this is not necessarily the same term as the superordinate term (for the term corresponding to the specific concepts). For example, from a conceptual point of view, *knife* and *scissors* are "cutting instruments". Nevertheless, it is not possible to naturally substitute *cutting instrument* for *knife* and *scissors*, as in:

(12) a. Use a knife and scissors to cut up the duck.
b.? If you don't have these cutting instruments, pull the duck apart.

There is no consistently used term for expressing the generic concept of *knife* or *scissors*. This can be *cutting instruments* as well as *instruments for cutting* or *cutting utensils*. Whether or not such a term exists varies among languages. For example, in Mandarin Chinese, the term *dāo* is fully accepted as a superordinate term to point to the Chinese equivalents of *knife* and *scissors*. In English, a term like *vegetable* is the superordinate of *carrot*, *tomato* or *cucumber* because it is consistently used for these previous words in texts.

This entails that choice of superordinate terms as lexical anaphors cannot be made at the conceptual level alone.

Moreover, superordinate terms can often be more easily used to lexicalize reference to a non-homegeneous set of elements than for reference to a single element or homogeneous set, as illustrated in (13) and (14):

(13) a. Put the carrots in to boiling water.
b. ? Remove the vegetables after 10 minutes.
(14) a. Throw the carrots, the leeks and the potatoes in to boiling water.
b. Remove the vegetables after 10 minutes.

However, the ease with which a superordinate can be used depends on the particular noun. For example, in French, *viande* [meat] can be subsituted for *bœuf* [beef] even in singular:

(15) a. Mettre le bœuf à cuire dans l'eau bouillante.
[Put the beef in the boiling water]
b. Retirer la viande au bout de 20 minutes.
[Remove the meat after 20 minutes]

This somewhat surprising phenomenon can be analysed with the help of the notion of basic level object proposed by Rosch *et al.* (1976). The importance of the basic level distinction for text generation has recently been shown by Reiter (1990). Rosch *et al.* demonstrated that the taxonomy of concepts could be organized using a structure with three levels: superordinate, basic and subordinate. They define the basic level as follows: "basic objects are the most inclusive categories whose members: (a) possess significant numbers of attributes in common, (b) have motor programs which are similar to one another, (c) have similar shapes, and (d) can be identified from averaged shapes of members of the class" (Rosch *et al.* 1976: 382)

It has been shown that lexemes corresponding to basic level objects seem to be the most natural terms to introduce referents already identified. For example, if one wants to refer to some *champignons de Paris* [button mushrooms], one would prefer to call them *champignons* [mushrooms], provided that there is no potential ambiguity with any other mushrooms. *Champignons de Paris* would seem too specific in this context and *vegetables* would seem too vague. This choice is not made randomly: *champignon* is the noun corresponding to the highest basic level concept to designate these objects. This would explain why in (15), one can refer to *bœuf* with the superordinate *viande*.

Nevertheless, the notion of basic level object does not always seem well suited to explain phenomena such as that observed in (15). For example, it seems that the concept "volaille" ["fowl"] fits perfectly the four criteria given by Rosch. But, *volaille* [fowl] does not seem a

natural French term for referring to a chicken, particularly in the sublanguage of recipes.

It is also problematic that the naming of basic level objects varies a great deal among languages. For example, in Mandarin Chinese, the most natural term to designate a knife when there is no ambiguity is the term *dāo* , which corresponds to "cutting instrument" in English. We could argue that conceptual representation differs with the mother tongue of the speaker (which is plausible, without joining the debate about language and thought) and that the lexicon reflects the conceptual views. Nevertheless, this position does not solve the problem of terms like *volaille*, a unnatural term for a basic level object.

It is significant that this position creates practical problems for text generation: if conceptual representation is reflected too closely in the choice of lexemes, this representation cannot be used as an interlingua for multilingual generation or machine translation.

In the light of this evidence, we have decided in favor of a strict theoretical separation between conceptual representation and lexical representation. We believe that an appropriate conceptual representation can be used for multilingual generation because it is a non linguistic generalization above specific lexical representations. We therefore distinguish the notion of basic level object, which belongs to cognitive science, from the notion of basic noun, which is a linguistic notion[6]. We consider "viande" and "volaille" to be basic level objects while only *viande* is a basic noun.

For lexical choice in text generation, we thus have to distinguish two very different processes:

- Superordination should be used to introduce a noun which points back to a set of different nouns. This is the case in {*carrots*, *leeks*, *cucumber*} --> *vegetables*. This process obeys a principle of economy.
- Basic denomination is used to introduce the most natural term for a given referent or a set of referents. This process obeys a principle of "naturalness": it introduces the most closely basic noun that corresponds to the concept to be lexicalized. Basic denomination is often used in texts like recipes: objects are first introduced with extreme precision and subsequently referred to with the basic term.

4. Knowledge sources for determining lexical anaphors

In the course of our work, we have proposed a series of algorithms for generating grammatical and lexical anaphora in procedural texts (Tutin 1992). Contrary to lexical anaphora, grammatical anaphora makes use of closed lexical classes (determiners, pronouns and a few special verbs) as well as ellipsis.

These algorithms are derived from an empirical study of French recipes, using a representative corpus of over 16,000 words. Recipes serve as a good prototype of procedural texts for assembling complex objects from parts. Even this modest corpus presents a wide variety of lexical and grammatical anaphora which are typical of assembly instructions.

We describe below some of the knowledge sources and organization needed to generate lexical and grammatical anaphora. For lack of space, however, we leave out the model of state change management (needed to describe recipe ingredients being mixed together and transformed (Kosseim 1992)), and the focus model used.

4.1 Input

We limit our scope to the linguistic part of generation; therefore, we assume that our input is the output of a text planner, which has already grouped actions into discourse structures as proposed by Grosz and Sidner (1986) and (Dale 1988). The input is thus a sequence of actions and states in which participants (ingredients, instruments and agent) are represented by indices.

4.2 Dictionary of concepts

The dictionary of concepts has been inspired by Nirenburg and Raskin 1987; concepts are mainly subdivided into actions or objects. We have added a category of properties, needed to describe relations between concepts (e.g., temporal limit) or attributes (e.g. *size*).

Relations between concepts are *isa*, *part-of* or *result*, the latter one useful in a domain where state changes are frequent. Thus, one can relate the action "cut" to the concept "piece" which is the result of "cut". The dictionary of concepts is not a copy of the language and there are concepts without any corresponding lexicalization. Taxonomic organization is functional and depends greatly on the field for which it has been established. In other words, our description of concepts has limited value outside the domain of recipes.

4.3 Dictionary of lexical entries

The representation of lexical entries is strongly influenced by the ECD (Mel'čuk & Polguère 1987, Mel'čuk *et al.* 1988). Two parts of the entry are particularly interesting for our topic: the semantic zone and the LF zone.

The semantic zone contains four types of information:
- The semantic field to which the lexeme belongs. For example, the verb *simmer* would have feature /cook/.
- The mass/count feature.
- The "basicness" feature, if the lexical item is a noun, indicates whether or not the noun is a basic noun.
- The key word(s) for which the lexeme can be a value. For example, for the lexeme *mixture*, it will be stated that it is the S_{res} of *mix*.

In the LF zone, we simply enumerate the values of the lexical item as a key word. For example, the entry for the verb *hacher* [chop] may contain, among many others, *hachis* ($S_{res}(hacher)$) and *hachoir* ($S_{med}(hacher)$).

5. Constraints for generating anaphors

We now turn to the constraints which apply to the choice of grammatical or lexical anaphors during text generation. Our aim here is to generate the most appropriate anaphor with respect to the textual context. To determine what is appropriate, we have used an empirical approach, rather than appeal to general principles such as Gricean conversational maxims (see Reiter 1990a & Dale 1988 for use of these notions for lexical choice in text generation). A detailed

6) Wierzbicka (1985) has shown in lexicographic descriptions that the names of (words for) basic level objects have special semantic properties..

examination of our corpus of cooking recipes has shown that anaphora is not governed so much by strict rules as by tendencies. Thus, in a given context, a set of possible anaphors can "compete" for selection. When choosing from multiple possibilities we favor the most "economical" anaphor, i.e., the one which conveys the least information[7].

Space limitations prevent a complete discussion of all factors required for an anaphor choice algorithm (see Tutin 1992). Here we give the most important constraints on choice among the principal anaphoric devices[8].

The selection of an anaphoric device has two stages:
- First, a choice is made among of grammatical devices (e.g. personal pronoun, verb complement ellipsis, coreferential definite NP, demonstrative NP).
- Then, if a lexical NP has been chosen, the correct lexical anaphor is determined.

5.1 Grammatical anaphora

The introduction of a given grammatical anaphor depends mainly on 4 kinds of parameters: a) the conceptual nature of referents, b) distance to antecedent and discourse structure, c) focalization and d) potential ambiguity.

We briefly review these different parameters for each type of grammatical anaphor: verbal complement ellipsis, personal pronoun, demonstrative NP, coreferential definite NP.

Verbal complement ellipsis as in the following example is very widespread in recipes, and characteristic of procedural instructions in general .

(16) Prepare the carrots, the celery and the asparagus. Cook Ø in the boiling water and take Ø out after 10 minutes.

Verbal complement ellipsis is generally used to designate a heterogeneous set of objects, contrary to personal pronouns. The distance from the antecedent can be quite far but focalization constraints, in particular global focus - defined as the subset of the most salient items - play a determining role for the production of this anaphor.

A personal pronoun must name an object or a set of **similar** objects. It is governed by very strong locality constraints (Hobbs 1978) and, as previously noted in the literature, personal pronouns often maintain the thematic continuity (Van Dijk & Kintsch 1983), i.e. pronoun is the local focus (what the clause is about) of both the previous and the current clauses. In fact, local focus generally supplies enough information for the hearer to correctly interpret the pronoun (as emphasized by Grosz, Joshi & Weinstein 1983), even if it is morphologically ambiguous.

Choice of a demonstrative NP does not depend on the conceptual nature of the referent, which may be either the local focus or the global focus. Its contrastive functions with respect to personal pronouns and definite NPs are rather complex. Since demonstratives are infrequent in our corpus, they are not treated further here.

For a definite NP, there is no conceptual restriction on the referent. A definite NP can be introduced at substantial distance from its textual antecedent, and typically does not occur in the following clause, especially if the antecedent was the local focus of its clause and there is no potential ambiguity[9] .

For each NP to be generated, potential ambiguity must be taken into account. This has to do with lexical choice. For example, choice of an ambiguous NP such as *le vin* [the wine] must be blocked if there is white wine and red wine in the context. The context in which the anaphoric NP must be distinctive depends on the anaphor chosen: it is the preceding sentence for demonstrative NP while, for definite NP, a larger context must be taken in account[10].

5.2 Lexical anaphora

We now turn to the constraints on choice of lexical anaphor. When the grammatical mechanism chosen for expressing anaphora involves a coreferential (definite or demonstrative) lexical NP, these constraints come into play to pick the most appropriate lexical form. The anaphoric lexical devices presented here for recipes constitute only a subset of those that could appear in the language as a whole. Nevertheless, we hypothesize that the conceptual and linguistic constraints governing their usage are generalizable to other kinds of text. Lexical anaphora differs significantly on this point from grammatical anaphora, whose constraints, like discourse structure or focalization, vary greatly according to the kind of text. Therefore, while a given kind of text might use only a subset of possible lexical anaphoric devices, these devices are governed by the same constraints in all kinds of texts. For example, typical result mention (*mix -> mixture*) is widespread in procedural texts but constraints governing them are the same in any kind of text. In contrast, it appears that the constraints governing usage of grammatical anaphoric devices, and even the devices themselves, are much more dependent on the variety of text.

Given that a lexical NP has been chosen, as the general type of anaphoric device, two kinds of constraints, conceptual and linguistic, apply to select the the specific kind(s) of lexical anaphora which may be used. In case of ambiguity, i.e. if the NP produced is not distinctive, additional processing will be required.

Conceptual constraints concern mainly:

- The state of the object . For example, an object whose state is being transformed by an action should be referenced via its resulting state.
- Groupings of objects: is the referent to be generated a set of identical objects, a heterogenous set, a homogenous set or a single element? A heterogenous set is composed of elements which just have no close

---

7) Anaphoric devices thus have a default (strict) order of priority for application.

8) We omit the realization constraints, such as the fact that certain verbs do not allow their complements to be pronominalized.

9) For example, the definite NP in the second clause is not very natural in French:

Marie a rencontré un charcutier. Le charcutier fait un très bon pâté.

[Marie met a porkbutcher. The porkbutcher makes very good pâté.]

10) For recipes, we use Dale's (1988) proposal to take the whole text as context, since it is usually short. This would of course not be satisfactory for longer texts,

generic concepts in common, such as, {"salt","knife","table"}.

Linguistic constraints involve mostly the lexical form and relative order of the coreferential NPs that have been lexicalized in the preceding text. Therefore, we do take advantage of referents already lexicalized in the previous context (which must be stacked for being available when lexicalizing).

The following properties are examined:
- The linguistic form of antecedent NP: is it a single noun, a compound noun or a complex NP?
- The existence of a lexico-semantic association for the antecedent like the generic term or the typical result (which can mostly be formalized through a LF).
- The "basicness" of the head word of the antecedent NP.

Ambiguity constraints are used to check if the lexicalization is not ambiguous.

If a unique object or a set of identical objects can not be lexicalized in a non ambiguous way, we lexicalize it the same way it has been first introduced in the text (Initial strict repetition). We use this ad hoc strategy because first mention of a referent is generally the most accurate. Of course, this would not always be the minimal distinguishing description (Dale 1988), but as Reiter (1990a) points out, determining a minimal distinguishing description may require overly complex processing.

In case of potential ambiguity for a set of heterogeneous objects, we use "complex coordination". With this process, we regroup first the first level superordinates and apply the other devices to the remaining list of objects[11].

Table 2 shows several important kinds of lexical anaphoric devices, with their associated conceptual, linguistic and non ambiguity conditions.

Lexical anaphor	Conceptual Properties	Linguistic Properties	Non ambiguity Constraints	Examples
Strict Repetition	Unique object or set of identical objects	Antecedent is a single noun (or fixed compound) and is a basic noun	No instance previously introduced has the same repetition	lapin --> lapin [rabbit]
Initial Strict Repetition	Unique object or set of identical objects	The other devices are ambiguous	No constraints	A small rabbit ... the rabbit --> the small rabbit
Partial Repetition	Unique object or set of identical objects	Antecedent is a not fixed compound (except "part-of" types) and the NP head is a basic noun	No previously introduced NP has the same partial repetition	petit lapin [small rabbit] --> lapin
Superordination	Set of objects having a close common generic concept	Nominal heads of antecedents have the same common superordinate. LF: Gener	No previously introduced NP has the same superordinate term	{carottes, poireaux, tomates} [carrots, leeks, tomatoes] --> légumes [vegetables]
Basic Denomination	Unique Object or set of identical objects	Nominal head of NP is not a basic noun	No previously introduced NP has the same basic denomination	petites girolles [small chanterelles] --> champignons [mushrooms]
Nominalization	Action	Antecedent verb can be nominalized or superordinate of antecedent verb can be nominalized. LFs: S_0 or S_0 o Gener	No constraints	faire cuire le poulet [cook the chicken] --> la cuisson du poulet [the cooking of the chicken]
Typical Result Mention	Object(s) having been affected by a strong transformation	There is a result noun for the actants having been affected by the tranformation. LF: S_{res}	No previously introduced NP has the same result mention	mélanger les patates$_i$ [mix the potatoes$_i$] --> le mélange$_i$ [the mixture$_i$]
Complex Coordination	Set of different objects which have no common generic concept	No element of the coordination is ambiguous	No constraints	{petit lapin, grosses chanterelles} [small rabbit, big chanterelles] --> le lapin et les champignons [le lapin et les champignons]

Table 2: Constraints governing the introduction of lexical anaphora

11) We choose here to apply superordination separately to each instance: we do not allow regroupings of elements for superordination or typical result mention because, as Kosseim has noticed, we would have to process all the subsets to generate correct lexicalizations.

Conclusion

In this paper, we have described some of the problems raised making lexical choice in textual context, in particular for coreferential lexical anaphora. We have showed that paradigmatic Lexical Functions are well suited for creating lexical coreferential links. We have also distinguished the selection of superordinate term, which is used to point back to a set of different words, from selection of basic denomination, which is used to name in the most natural way a concept already introduced by a previous noun.

A series of constraints has been formulated which can be implemented in an algorithm for selecting among natural grammatical and lexical anaphors in procedural texts. Most of these algorithms have been implemented by Kosseim 1992. The generator uses Prolog and specifically Definite Clause Grammar (DCG) to produce text.

We find that determination of grammatical anaphora is more dependent on the genre and sublanguage than is lexical anaphora, which appears governed by fairly general constraints. However, more work needs to be done to check these results in other procedural texts, and then more broadly in less similar text types. Also, it would be interesting to see to what extent anaphoric expressions share common constraints with deictic expressions for which the context of interpretation is not the previous text, but the extra-linguistic context.

Acknowledgements

We would like to thank Guy Lapalme, Igor Mel'čuk, Alain Polguère, Marga Alonso Ramos and Xiaobo Ren for fruitful discussions and helpful suggestions. Special thanks to Leïla Kosseim, who collaborated in this research and with whom we shared many interesting discussions. The work reported in this paper was supported by a Government of Canada Award.

References

Bateman J. A. & L.Wanner (1990). Lexical Cooccurrence Relations in Text Generation, in *Proceedings of the Fifth International Workshop on Natural Language Generation*, Dawson, Pennsylvania, 31-38.

Dale, R. (1988). *Generating Referring Expressions in a Domain of Objects and Processes*, Ph..D. thesis, University of Edinburgh.

Décary M. & G. Lapalme (1990). An Editor for the Explanatory Dictionary of Contemporary French (DECFC), *Computational Linguistics*, 16, 3, 145-154.

Grosz B.J., Joshi A., Weinstein S. (1983). Providing a Unified Account of Definite Noun Phrases in Discourse, in *Proceedings of the 21st Annual Meeting of the ACL*, MIT, Cambridge, Mass., 15-17 june, 1983, 44-49.

Grosz B.J. & C. Sidner (1986). Attention, Intentions and the Structure of Discourse, *Computational Linguistics*, 12, 175-204.

Halliday M.A.K & R. Hasan (1976). *Cohesion in English* London, Longman.

Heid U. & S. Raab (1989). Collocations in Multilingual Generation, in *Proceedings of EACL*, 130-136.

Hobbs J. (1978). Resolving Pronoun References, *Lingua*, 44, 311-338.

Iordanskaja L., R. Kittredge & A. Polguère (1991). Lexical Selection and Paraphrase in a Meaning-Text Generation Model in C.L. Paris, W. R. Swartout & W.C. Mann eds.,*Natural Language Generation in Artificial Intelligence and Computational Linguistics*, 293-312.

Kosseim L. (1992). *Génération automatique de procédés cohésifs dans les recettes de cuisine*, M.Sc. thesis, Département d'informatique et de recherche opérationnelle, Université de Montréal.

McDonald D. (1991). On the Place of Words In the Generation Process in C.L. Paris, W. R. Swartout & W.C. Mann eds.,*Natural Language Generation in Artificial Intelligence and Computational Linguistics*, 229-247.

Mel'čuk I.A. *et al.* (1988a). *Dictionnaire Explicatif et Combinatoire du Français Contemporain. Recherches Lexico-sémantiques. II.* Montréal, Presses de l'Université de Montréal.

Mel'čuk I. (1988b). Paraphrase et lexique dans la Théorie Sens-Texte, *Cahiers de Lexicologie* LII, 5-50 et LIII, 5-53.

Mel'čuk I. & A. Polguère (1987) A Formal Lexicon in the Meaning-Text Theory (or how to do Lexica with Words), *Computational Linguistics*, 13, 3-4, 1987, 261-275.

Mel'čuk I. & Zholkovsky, D (1970). Sur la synthèse sémantique, *T.A. Informations*, 2, 1-85.

Nirenburg S. & I. Nirenburg (1988). A Framework for Lexical Selection in Natural Language Generation, in *Proceedings of COLING 88*, Budapest, 471-475.

Nirenburg S. & V. Raskin (1987). The Subworld Concept Lexicon and the Lexicon Management System, *Computational Linguistics*, 13,3-4, 276-289.

Reiter E.B. (1990a). Generating Appropriate Natural Language Object Descriptions, Ph.D. Thesis., Harvard University.

Reiter E.B. (1990b) A New Model for Lexical Choice for Open-Class Words, in *Proceedings of the Fifth International Workshop on Natural Language Generation*, Linden Hall Conference Center, Dawson, Pennsylvania, 23-30.

Rosch E., C. B. Mervis, W. D. Wayne, D. M. Johnson & P. Boyes-Braen (1976). Basic Objects in Natural Categories, *Cognitive Psychology* 8, 382-439.

Sidner C. (1983). Focusing in the Comprehension of Definite Anaphora, in M. Brady & R. Berwick eds.,*Computational Models of Discourse*, Cambridge (UK.), Cambridge University Press, 267-330.

Tutin A. (1992) *Etude des anaphores grammaticales et lexicales dans la perspective de la génération automatique dans des textes de procédures*, Ph.D. Thesis, Département de linguistique, Université de Montréal.

Wierzbicka A. (1985). *Lexicography and conceptual analysis*, Ann Arbor (Mich.), Karoma Publishers inc.