Collocations in Multilingual Generation

Ulrich Heid, Sybille Raab
Universität Stuttgart, Projekt Polygloss
Institut für maschinelle Sprachverarbeitung
Keplerstrasse 17
D-7000 Stuttgart 1, West Germany

Abstract

We present a proposal for the structuring of collocation knowledge [1] in the lexicon of a multilingual generation system and show to what extent it can be used in the process of lexical selection. This proposal is part of Polygloss, a new research project on multilingual generation, and it has been inspired by work carried out in the SEMSYN project (see e.g. [RÖSNER 1988][2]). The descriptive approach presented in this proposal is based on a combination of results from recent lexicographical research and the application of Meaning-Text-Theory (MTT) (see e.g. [MEL'ČUK et al. 1981], [MEL'ČUK et al. 1984]). We first outline the overall structure of the dictionary system that is needed by a multilingual generator; section 2 gives an overview of the results of lexicographical work on collocations and compares them with "lexical functions" as used in Meaning-Text-Theory. Section 3 shows how we intend to integrate collocations in the generation dictionary and how "lexical functions" can b used in generation.

1 Lexical knowledge for multilingual generation

Within a multilingual generation system, seems necessary to keep the dictionary a modular as possible, separating informatic that pertains to different levels of linguist description[3]. We assume that the system's le: ical knowledge is stored in the following typ of "specialized dictionaries":

- semantic: inventory of possible lexicaliz tions of a concept in a given language;
- syntactic: one inventory of realizatic classes per language, providing inform tion about number, type and realizatic of the arguments of a given lexeme;
- morphological: one inventory of infle tional classes per language.

Since none of these levels of decsriptic is completely independent, the dictionari should be linked to each other by means cross-references and reference to class men bership. Templates and mechanisms allov ing for explicit inheritance of shared prope ties, e.g. redundancy rules, will be used withi

[1] We use the term "collocation" in the sense of [HAUSMANN 1985] referring to constraints on the cooccurrence of two lexeme words; the two elements are not completely freely combined, but one of them semantically determines the other one. Examples are for instance *solve a problem, turn dark, expose someone to a risk*, etc. For a more detailed definition see section 2.

[2] Research reported in this paper is supported by the German Bundesministerium für Forschung und Technologie, BMFT, under grant No. 08 B 3116 3. The views and conclusions contained herein are those of the authors and should not be interpreted as positions of the project as a whole.

[3] For more details on the dictionary structure s [HEID/MOMMA 1989].

each of the layers. These mechanisms give access to the knowledge about the linguistic "behaviour" of lexemes needed in the process of lexicalization[4].

2 Approaches to the description of collocations

2.1 Contributions from lexicography

The tradition of British Contextualism[5] defines collocations on the basis of statistical assumptions about the probability of the coocurence of two lexemes. Particularly frequent combinations of lexical units are regarded as collocations.

A more detailed definition can be found in the work of Franz Josef Hausmann (1985:119):

> "One partner determines, another is determined. In other words: collocations have a basis and a cooccurring collocate."[6]

This determination manifests itself in so far as a given basis does not allow all of the collocates that would be possible according to general semantic coocurrence conditions, but only a certain subset: so in French, *retenir son admiration, retenir sa haine, sa joie* are possible, but **retenir son désespoir* is not.

The choice of collocates depends strongly on the lexeme that has been chosen as the basis; knowledge about possible collocations can be only partly derived from knowledge about general semantic properties of lexemes. Therefore general cooccurrence rules or selectional restrictions (e.g. using semantic markers) are not adequate for the choice of collocates in the process of lexicalization.

These considerations lead to two proposals for the structuring of the lexical knowledge used in a generator:

- Heuristic for the lexicalization process:

  > "First the basis is lexicalized, then the collocate, depending on which lexeme has been chosen as the basis."

- Knowledge about the possibility of combining lexemes in collocations should be stored in the lexicalization dictionary (where lexicalization candidates for concepts are provided), and specifically in the entries for the bases.

The following table shows in terms of categories[7] what can be a possible collocate for a particular basis[8]:

basis	possible collocates
noun	noun, verb , adjective
verb	adverb
adjective	adverb

[4] Possibly including classifications according to semantically motivated lexeme classes and a modelling of paradigmatic relations between lexemes, such as hyponymy or synonymy.

[5] The term "collocation" was introduced into linguistic discussion by John R. Firth (1951:94).

[6] Translation by the authors. We use the terms *basis* and *collocate* in the sense of [HAUSMANN 1985]; HAUSMANN's original terms are *Basis* and *Kollokator*.

[7] Unlike British Contextualism (cf. the recent [SINCLAIR 1987]) we assume that bases and collocates are of one of the following categories: noun, verb, adjective or adverb.

[8] For substantive-verb-collocations, the classification as basis and collocate is opposed to the usual syntactic description according to head and modifier; this has consequences for the lexicalization process: while it is usually possible to first lexicalize the heads of phrases, then the modifiers (e.g. $\text{substantive}_{head,basis} < \text{adjective}_{modifier,collocate}$, the choice of verbs depends on their nominal complements (which are modifiers, but which have to be considered as bases of collocations). This means that nouns have to be lexicalized before verbs, e.g. *Pläne schmieden*, but not **gute Vorsätze schmieden*).

2.2 Lexical functions of the Meaning-Text-Theory as a tool for the description of collocations

In MTT, developed by Mel'čuk and coworkers, there exist about 60 "lexical functions" which describe regular dependencies between lexical units of a language. In MTT, lexical functions are understood as cross-linguistically constant operators (f), whose application to a lexeme ("keyword", L) yields other lexemes (v). Mel'čuk (1984:6), (1988:31f) uses the following notation:

$$f(L) = v$$

The result of the application of a lexical function to a given lexeme can be another "one-word" lexeme, or a collocation, an idiom or even an interjection.

The parallelism between the collocation definition used in this paper and the notion of lexical function is that both start from the principle that collocates depend upon the respective bases (in MTT, v is a function of L). Therefore lexical functions seem to be a useful device for the description of collocations in a generation lexicon.

In the following, we only consider lexical functions which, when applied to a lexeme word, yield collocations[9]; Table 1 gives some examples of such lexical functions, together with a definitional gloss, taken from [STEELE/MEYER 1988][10]:

3 Generating Collocations

We propose that every lexeme entry in the lexicalization dictionary contains slots for lexica functions, whose fillers are possible collocates within a slot/filler-notation as the one usec in Polygloss, a (partial) lexical entry, e.g. fo *problem*, could be represented in the followin way:

```
(problem
    (...)
    (caus func (create, pose))
    (real (solve, ...))
    (...))
```

It might be possible to predict the type of lexical functions applicable to a given lexeme from its membership in a semantic class Syntactic properties of bases and collocates ar accessible through reference to the realizatio lexicon.

[MEL'ČUK/POLGUÈRE 1987]:271f themselves stress the advantage of describing collocations with lexical functions withi language generation and machine translation they give the example of OPER (*QUESTION*) realized as

- English *ask a question*,
- French *poser une question*,
- Spanish *hacer una pregunta* and
- Russian *zadat' vopros*

respectively[11].

3.1 Lexicon structure and possibl generalizations

On the basis of the analysis of some entrie in [MEL'ČUK et al. 1984] and of material w

[9] It should be investigated to what extent the category of v is predictable for every f, according to the category of L. For instance, f's of group 1 and 2 specified in the table below, applied to nouns, yield substantive+verb-collocations, those of groups 3 and 4 yield substantive+adjective-collocations, and those of groups 5 and 6 return substantive+substantive-collocations.

[10] Lexical functions of group 2, normally occur together with those from 1; ABLE only occurs in combination with other lexical functions.

[11] Here *QUESTION* refers to a concept that stand for the language-specific items.

	Lexical Functions	Meaning	Examples
1.	OPER, FUNC, LABOR, REAL, FACT, LABREAL	occurrence realization	OPER(*attention*) = *pay* REAL(*promise*) = *keep*
2.	PROX, INCEP CONT, FIN CAUS, PERM LIQU	phases phase + [CAUSE]	INCEP OPER(*form*) = *take* CAUS FUNC(*problem*) = *create, pose*
3.	MAGN, POS, VER	(high) degree	MAGN(*eater*) = *big, hearty* VER(*praise*) = *merited*
4.	ABLE, QUAL	ability	ABLE$_2$(*writing*) = *readable*
5.	MULT, SING	count ↔ mass	MULT(*goose*) = *gaggle*
6.	GERM, CULM	germ, culmination	CULM(*joy*) = *height*

Table 1: Examples of lexical functions used for the description of collocations

have analysed within Polygloss[12], it seems possible to generalize over some regularities in collocation formation for members of semantically homogenous lexeme classes.
An example: the following default assumptions can be made for nouns expressing information handled by a computer (we assume semantic classes *I-NOUNS$_G$* and *I-NOUNS$_F$* for German and French respectively):

- *I-NOUNS$_G$* = { *Datei, Information, Nachrichten, Verzeichnis* }
- *I-NOUNS$_F$* = { *fichier, information, messages, répertoire* }

LIQU FUNC$_0$(*I-NOUNS$_G$*) = *löschen*
LIQU FUNC$_0$(*I-NOUNS$_F$*) = *supprimer*

Some exceptions, however, have to be stated explicitly, as illustrated by the example of French nouns expressing personal attitudes, treated in [MEL'ČUK et al. 1984]:
PA* = { *admiration, colère, désespoir, enthousiasme, envie, étonnement, haine, joie, mépris, respect* }

OPER$_1$(*PA*) = *ressentir* ⟨ SUBJ OBJ
(OBJ PRED) ϵ *PA*

Exception:
OPER$_1$(*admiration*) = *nourrir* ⟨SUBJ OBJ⟩,
(OBJ PRED) =
"*admiration*"

OPER$_1$(*haine*) = *nourrir* ⟨SUBJ OBJ⟩,
(OBJ PRED) = "*haine*"

3.2 The generation of paraphrases

One of the aims in the development of the "how-to-say"-component of a generation system is to ensure that variants (i.e. true paraphrases) can be generated for one and the same semantic structure.

This involves two types of knowledge: more 'static' knowledge about interchangeability of realization variants (synonymous items, information about paraphrase relations between certain constructions or between collocations) and more 'procedural' knowledge about heuristics guiding the choice between candidates. The 'static' knowledge should be represented declaratively. It can be divided into information about syntactic variants (e.g. participle form vs. relative clause) and information about lexicalization variants. In

[12] Manuals for PC-Networks that have been provided in machine-readable form in German and French by IBM; cf. [RAAB 1988].

[MEL'ČUK 1988]:38-41 rules are stated, which express paraphrase relations between certain types of collocations. Ideally these rules can be set up for pairs of lexical functions, without consideration of concrete lexemes. Examples are:

- Jean *s'est mis en colère* contre Paul
  (=INCEP OPER$_1$)
  John got angry with Paul
  ⟷
  Paul *s'est attiré la colère de* Jean.
  (=INCEP OPER$_2$)
  Paul angered John.
- Jean *s'est pris d'* enthousiasme pour cette découverte.
  (=OPER)
  John got enthusiastic about this discovery.
  ⟷
  (A cause de cette découverte)
  l'enthousiasme s'est emparé de Jean.
  (=FUNC)
  John was enthused by this discovery.

Within a generation system, such descriptions can be used to state paraphrase relations between collocational lexicalization candidates. The choice between candidates depends on parameters, amongst which the following ones seem to be essential:

- syntactic "behaviour" of the lexemes building up a collocation[13]
  - in relation to roles in the frame structure to be realized;
  - in relation to the thematic structure of the intended utterance;
- markedness of lexemes (e.g. registers style);
- general heuristics for text generation (e.g "avoid repetition", "avoid deep embedding" etc.)

In the following, we give an example fo the lexicalization possibilities that can be described with the proposed device:
given the following (rudimentary) semanti representation[14]:

mental process : *BE-HAPPY*
:BEARER *PIERRE*
:CAUSE *NEWS*,

there should be available the following information about collocations with *joie* as basis[15]:

CAUS FUNC(*joie*)	=	*causer la joie de qn,* *causer de la joie chez qn*
CAUS OPER(*joie*)	=	*réjouir qn,* *mettre qn en joie* *remplir qn de joie*
INCEP FUNC(*joie*)	=	*la joie s'empare de qn* *la joie saisit qn,* *la joie naît dans le coeur de qn*
INCEP OPER(*joie*)	=	*qn se met en joie*

The choice between INCEP and CAUSE depends on whether (and how) the causality is t be expressed. The choice between INCEP OPE and INCEP FUNC depends on whether the realaization of *PIERRE* or of *NEWS* should become the subject.

[13] We plan to investigate to what extent it is possible to describe the syntactic form of certain collocations with general rules. This is possible e.g. for OPER, FUNC, LABOR, i.e. for lexical functions yielding collocations of the type of "Funktionsverbgefüge":

OPER(L)	⟶	verb ⟨ SUBJ OBJ ... ⟩ (OBJ PRED) = L
FUNC(L)	⟶	verb ⟨ SUBJ ... ⟩ (SUBJ PRED) = L
LABOR(L)	⟶	verb ⟨ SUBJ OBJ Y ⟩ (Y PRED) = L

[14] *mental process* is meant to be a concept typ :BEARER and :CAUSE are semantic relations; *BE HAPPY*, *PIERRE* and *NEWS* are concepts.

[15] In simplified notation. The first two examples ai roughly equivalent to English make someone happy, fi someone with joy, the latter ones to *to please someon*

Here constraints caused by the syntax of the utterance to be generated play an important role: in a relative clause e.g. the antecedent has already been introduced. This fact limits the choice:

- - *... et alors cette nouvelle arriva, qui ...*
  - *causa la joie de Pierre* (= CAUS FUNC)
  - mit *Pierre en joie* (= CAUS FUNC)
- *... et alors Marie envoya cette nouvelle à Pierre, qui ...*
  - *se réjouit* (= CAUS FUNC)
  - *se* mit *en joie* (= CAUS FUNC)

This example shows that the heuristic "lexicalize bases first, then collocates" interacts with constraints stemming e.g. from syntax; these constraints can also be produced by a text structuring component (decisions about topic, thematic order etc.). The modular design of the lexicon supports generation of variants by giving access to all information needed at the appropriate choicepoints.

4 Conclusion and directions for future work

We propose a method for the description of knowledge about collocations in the dictionary of a multilingual generation system. Advantages for text generation result from the application of MTT's lexical functions and the formulation of the heuristic discussed above.

In the generation literature, the generation of collocations is regarded as a problem (cf. [MATTHIESSEN 1988]). The only system we know of, in which attempts have been made to bring it to a solution, is DIOGENES, a knowledge based generation system under development at Carnegie Mellon University[16]. Our approach differs from NIRENBURG's in that it introduces the distinction between basis and collocate. This leads to differences in the lexicalization strategy: within DIOGENES, heads are lexicalized before modifiers, irrespective of word classes, cf. [NIRENBURG/NIRENBURG 1988].; we have come up with data that seems to favour the distinction between basis and collocate.

Further contrastive descriptive work will be the basis for a prototypical implementation within Polygloss. With respect to lexical functions, some questions related to defaults (e.g. syntactic realization defaults, inheritance of collocational properties within lexem classes etc.) should be investigated in more detail.

4.1 Acknowledgements

We would like to thank Sergei Nirenburg and our collegues at the IMS for the fruitful discussions in this paper. All remaining errors are of course our own.

References

[FIRTH 1951] John Rupert Firth: "Modes of Meaning." (1951) in: *Papers in Linguistics 1934-51.* (London) 1957 (SS.190-215)

[HAUSMANN 1985] Franz Josef Hausmann : "Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels." in: Henning Bergenholtz / Joachim Mugdan (Eds.): *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch.* 1985: 118-129 [= Lexicographica. Series Maior 3]

[HEID/MOMMA 1989] Ulrich Heid, Stefan Momma: "Layered Lexicons for Gen-

[16] For a general overview of DIOGENES, see [NIRENBURG et al. 1988]. Questions of lexicalization and of the treatment of collocations are treated in [NIRENBURG 1988], [NIRENBURG et al. 1988], [NIRENBURG/NIRENBURG 1988].

eration", internal paper, University of Stuttgart, IMS, 1989

[MATTHIESSEN 1988]
Christian Matthiessen: "Lexicogrammatical Choices in Natural Language Generation", ms., paper presented at the Catalina Workshop on Natural Language Generation, (Los Angeles), June 1988

[MEL'ČUK 1988] Igor A. Mel'čuk: "Paraphrase et lexique dans la théorie linguistique Sens-Texte." in: *Lexique* 6, Lexique et paraphrase. Lille 1988: 13-54

[MEL'ČUK et al. 1981] Igor A. Mel'čuk et al.: "Un nouveau type de dictionnaire: le dictionnaire explicatif et combinatoire du français contemporain (six entrées de dictionnaire)." in: *Cahiers de Lexicologie* (28) 1981-I: 3-34

[MEL'ČUK et al. 1984] Igor A. Mel'čuk et al.: *Dictionnaire explicatif et combinatoire du francais contemporain. Recherches Lexico-SΘmantiques. (I)*, Montréal 1984

[MEL'ČUK/POLGUÈRE 1987] Igor A. Mel'čuk, Alain Polguère: "A Formal Lexicon in the Meaning-Text Theory (or how to do Lexica with Words)." in: *Computational Linguistics* 13 3-4 1987: 261-275

[NIRENBURG 1988] Sergei Nirenburg: "Lexical selection in a blackboard-based generation system." Paper presented at the *Catalina Workshop on NL generation*, Los Angeles 1988, ms.

[NIRENBURG et al. 1988] Sergei Nirenburg et al.: "DIOGENES-88, CMU-CMT-88-107." Pittsburgh: CMU, 1988, ms.

[NIRENBURG et al. 1988] Sergei Nirenburg et al.: "Lexical Realization in Natural Language Generation." in : *Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages. Pittsburgh, Pennsylvania June 12 - 14, 1988, Proceedings*, 1988

[NIRENBURG/NIRENBURG 1988] Sergei Nirenburg, Irene Nirenburg: "Choosing Word carefully", (Pittsburgh, Pa.: ICMT, Carnegie-Mellon University), 1988, internal paper.

[RAAB 1988] Sybille Raab: *Zur Beschreibung fachsprachlicher Kollokationen*, ms., University of Stuttgart, 1988

[RÖSNER 1988] Dietmar Rösner: "The SEMSYN generation system", in: Proceedings of ACL-applied, Austin, Texas, February 1988, 1988

[SINCLAIR 1987] John McH Sinclair: "Collocation. A progress report." in: Ross Steele / Terry Threadgold (Eds.): *Language Topics. Essays in honour of Michael Halliday.* (Amsterdam/Philadelphia) 1987, vol. 2.: 319-331

[STEELE/MEYER 1988] James Steele, Ingrid Meyer: "Lexical Functions in the Explanatory Combinatorial Dictionary : Kinds and Definitions." Internal paper, Université de Montréal, 1988