

# A Constraint-based Representation Scheme of Collocational Structures

Dirk Heylen† André Schenk‡ Marc Verhagen‡

(heylen@let.ruu.nl) (schenk@let.ruu.nl) (verhm@sx.ac.uk)

† OTS – Trans 10 – 3512 JK Utrecht (NL)

‡ CLMT Essex University – Wivenhoe Park – C04 3SQ Colchester (UK)

## 1 Introduction

The main aim of the ET-10/75 project, 'Collocations and the Lexicalisation of Semantic Operations'<sup>1</sup>, is to evaluate the use of Mel'čuk's analysis of collocations in terms of *lexical functions*<sup>2</sup>, as an interlingua device in a machine translation system. In this poster we only point out the main design issues involved in encoding the essential parts of this analysis in an HPSG-style framework. Other aspects of the project, such as the definition of collocations, or the evaluation of the analysis as such, will not be dealt with. The linguistic analysis of collocations (as far as it is dealt with in this poster) involves two major issues.

- Coding the Mel'čukian analysis in an HPSG-like grammar.
- Designing the architecture of the lexicon.

## 2 Analysis

Examples of the collocational structures we have in mind are *strong criticism* (as an adjective-noun combination) and *give a demonstration* (as a verb-noun combination). The essential characteristics we want to account for in our analysis are the following.

- The collocational cooccurrence restriction.
- The dependence of the meaning of the collocate on the combination.
- The relation between the 'collocate occurrence' of a word and its free occurrence.

Following Mel'čuk, the restricted choice of the particular adjective or verb is represented in the dictionary by means of lexical functions. The dictionary entry for *criticism* contains a section in which its 'collocates' are listed, indexed by a lexical function. In this case, we would find:  $\text{Magn}(\text{criticism}) = \text{strong}$ . This points out that you can use *strong* when you want to express  $\text{Magn}(\text{criticism})$ . In our interpretation of these lexical functions, we take them to be general semantic operations, representing the major semantic contribution of the collocate. So we say that *strong* means  $\text{Magn}$  (= very, intense ...) in the context of *criticism*.

As far as translation is concerned, we no longer need a transfer rule mapping 'strong criticism' on

<sup>1</sup>We would hereby like to acknowledge the financial support by the Commission of the European Community, Association Suissetra (Geneva) and Oxford University Press.

<sup>2</sup>See for instance [Mel'čuk and Žolkovsky, 1988].

'scherpe kritiek' (lit. sharp criticism) because if both are analyzed as the interlingual  $\text{Magn}$ , we can leave the correct choice of adjective to the monolingual components.

The HPSG grammars we are using, closely resemble the proposals in [Pollard and Sag, 1987]. As far as the coding of the lexical functions is concerned, we have simply interpreted these as *relation names*.

## 3 Representation

Besides the use of a Mel'čukian analysis and the choice of an HPSG-style grammar, one of the design criteria that has guided the representation is to be 'minimally redundant'. This has led to the following solution.

- The collocate is only specified by a partial lexical entry, which is a *generalization* of its 'free-occurring' counterpart (its meaning is left unspecified, for instance).
- In the lexicon, it is specified on the base which collocates it *selects* for a particular lexical function. The remainder of the collocation specific information about the collocate is specified here as well.

The following is part of the entry for *criticism* to illustrate the latter issue. The COL field provides the information that to express  $\text{Magn}(\text{criticism})$ , it can combine with *strong*. The reference \$strong is to the base of the collocate entry in the dictionary.

$$\begin{array}{l} \text{\$criticism:} \\ \text{PHONcriticism} \\ \left[ \begin{array}{l} \text{SEM} \left[ \text{CONT|IND} \left[ \begin{array}{l} \text{VAR[1]} \\ \text{REST} \left[ \text{RELNcriticism'} \\ \text{INST[1]} \end{array} \right] \end{array} \right] \right] \\ \text{COL} \\ \{ \text{\$strong} \left[ \text{CONT|IND} \left[ \begin{array}{l} \text{VAR[1]} \\ \text{REST} \left[ \text{RELN MAGN'} \\ \text{INST[1]} \end{array} \right] \end{array} \right] \right] \} \end{array} \right] \end{array}$$

## References

- [Heylen *et al.*, 1993] Dirk Heylen, André Schenk, and Marc Verhagen. A unification-based representation scheme of collocational structures. to appear, 1993.
- [Mel'čuk and Žolkovsky, 1988] I.A. Mel'čuk and A.K. Žolkovsky. The explanatory combinatorial dictionary. In M. Evens, editor, *Relational Models in the Lexicon*, pages 41–74. CUP, Cambridge, 1988.
- [Pollard and Sag, 1987] Carl Pollard and Ivan Sag. *Information Based Syntax and Semantics*. CSLI, Stanford, 1987.