Ulrich Heid
Institut für maschinelle Sprachverarbeitung, Universität Stuttgart

On Ways Words Work Together – Topics in Lexical Combinatorics

1. Introduction

1.1 Broad areas of combinatory phenomena

The domain of lexical combinatorics has received much interest over the last years, in syntax, lexical semantics and lexicology, but also in lexicography, terminology, terminography and in Natural Language Processing (NLP). If the field of combinatorics can maybe trivially be defined by the fact that it deals with syntagmatic combination phenomena involving two or more lexemes, it is much harder to come up with any reasonable internal subdivision of the field.
Phenomena which are usually described as belonging to the domain of combinatorics include, among others:

- *Selectional properties of lexical items*: for example, the English verb *to grow* has, broadly speaking, two French equivalents, *pousser* and *grandir*. And most dictionaries would state that *pousser* is preferred if the subject noun denotes a plant, *grandir* if it denotes a human being.[1] The classical example of German *essen* ⇔ *fressen*, for English *to eat* (depending on the distinction between human being and animal) is another instance of this phenomenon.
- *Collocations*: according to many linguists and lexicographers,[2] collocations are combinations of exactly two lexemes (of category noun, verb, adjective or adverb), realizing two concepts, where the choice of one of them depends on (or: is restricted by) the other. Typical examples which are often cited are FR *un célibataire endurci*, DE *ein eingefleischter Junggeselle*, EN *pay attention*, FR *pousser un cri*, etc. Usually, some sort of determination relation between the two items can be found.[3]

  Other lexicographers and NLP researchers have a wider notion of collocation which subsumes any kind of combination of two words, as it occurs (adjacently) in a text. Such a wider view is not uncommon in work on statistical tools (where e.g. also the combination with closed class items may be regarded as a collocation, and where frequency, e.g. of co–occurrence, is the main definition criterion).

  Much of the discussion in this conference will be devoted to collocations; this is one of the reasons why we have chosen to discuss

collocations in some more detail, in this paper, taking them as a paradigmatic example of some of the research topics in the linguistic and lexicographic description of combinatory phenomena.

- *Idioms*: the common view on idioms is that they are multiword expressions (more than two items) which have an "en bloc–meaning" opaque with respect to the usual meaning of the words making up the combination. In examples like DE *das Kind mit dem Bade ausschütten*, we do not say anything about a child or a bath, somebody who FR *a(voir) une araignée au plafond* may also have other trouble than just with a spider.

As soon as we look at data from text corpora, cases come up where it is not easy to determine clearly whether to treat a given item as idiomatic or as collocational: DE *eine Frage stellen* is usually classified and described as a collocation (a support verb construction almost synonymous with DE *fragen*), whereas DE *in Frage stellen* is less clear: should it be treated as an idiom roughly equivalent to DE *anzweifeln* or as a collocation?

1.2 Structure of this paper

The purpose of this paper is to give an overview of some research topics in the field of lexical combinatorics. This includes a presentation of the main approaches, methods and strands of research, as of open issues and lines to be followed, in paricular those discussed at the Euralex–94 conference. Such an overview is bound to be partial, in both senses of the word: it is impossible to select all (and only the) relevant topics, and the selection is of course biased towards the preferences of the author.

Nevertheless, selecting collocations as a prototypical phenomenon seems to make sense from a more general point of view as well: collocational phenomena are central to lexicographers, corpus linguists and terminologists; evidence: the sheer number of papers on this topic submitted to the Euralex–94 conference. Moreover, the description and lexicographic modeling and representation of collocations is not at all an easy task: a few properties of collocations are well–known and easily reproducible, but others are controversial or not easy to consistently verify on data.

The problems which need to be addressed and which will be to some extent discussed in this paper fall into the following areas:

- defining the notion of collocation, delimiting it with respect to other combinatory phenomena and identifying criteria allowing to operationalize to some extent the definitions;
- describing syntactic, semantic and pragmatic properties of collocations and other combinatory phenomena, both within descriptive linguistic and lexicographic work (the latter including in addition to linguistic description also issues of the presentation of the descriptive results);

- getting, by means of computational tools for lexical acquisition, material potentially relevant for collocational description: techniques, methods and tools for extracting collocation candidates from texts;
- representing and using collocational information, for example in translation, both human and computer-aided or automatic.

These topics span a range of activities of (computational) linguists and (computational) lexicographers and terminologists: definition, description and lexicographic presentation of collocations, as well as their (semi-)automatic acquisition and use in human and computational applications of lexical knowledge sources.

We have chosen to comment on these topics in the following order:

- Definitional and descriptive problems, as treated in linguistic work on lexical combinatorics will be discussed, along with syntactic, semantic and pragmatic properties of collocations, in Section 2. This allows us to better capture the phenomenon we deal with, from different points of views.
- On this basis, we will deal with the lexicographic and terminographic treatment of collocations, including aspects of the presentation of descriptive results in dictionaries, in Section 3.
- The acquisition of collocationally relevant information from textual corpora, as well as the use of collocation knowledge in translation dictionaries will be the topic of Section 4.

We will illustrate some of the statements made in this paper with examples from dictionaries. The aim of this paper is not to support one given approach or to argue for a given method or tool for the acquisition or description of collocations: the examples have been chosen for their illustrative character, and an attempt has been made to cover several approaches.

2. Properties of combinatory phenomena – the case of collocations

2.1 Data and a first interpretation

The intuition about collocations is that they are combinations of two lexemes, not necessarily textually adjacent ones. To these two lexemes correspond two concepts. In certain collocations, we can find a regular semantic interrelationship between the two components which is close to a determination relation (collocations are "polar" in Hausmann's terms).

An essential property of collocations seems to be their perception by native speakers of a language as frequent, recurrent, conventionalized building blocks of the lexicon: "déjà-vu", as Hausmann says. The combination of exactly the two items appearing in the collocation is lexically determined; it is often not predictable; but native speakers are quite good at

identifying non–collocational combinations in other people's texts, and they feel that non–collocational texts are not fluent, not elegant or just not the "usual way" how one would express a given idea.

Collocations occur in both general language and sublanguage. The table in Figure 1 contains a few examples from English, French and German. The sublanguage examples may be felt to be different in nature from those given for general language: we will come back to this later (see Section 2.3.3).

language	general language	sublanguage
English	*pay attention,* *want sth. badly* *merited praise* *closely related*	*stop the conveyor* *overlaying rock* *expensive in labour*
French	*opérer un choix* *une déception amère* *éperdument amoureux*	*créer un fichier* *élution greduée* *ressources renouvelables*
German	*eine Vereinbarung treffen* *starker Raucher* *tief beeindruckt* *(jmdn) hart treffen*	*eine Forderung abtreten* *Abwasser einleiten* *anstehende Kohle* *Dateien abgleichen*

Figure 1. A few examples of general and sublanguage collocations

A number of criteria have been discussed in the literature to distinguish collocations from free combinations on the one hand and from idioms on the other, or, rather, to arrange examples of certain types somewhere on the scale between these two extremes. These criteria involve the syntactic, semantic and pragmatic description of lexemes.

2.2 Syntactic properties

2.2.1 Combinatory phenomena vs. phrase structure

Most combinatory phenomena follow the rules of syntax; no particular syntactic rules are necessary to describe combinatory phenomena. But not all of them make up constituents.

Selectional phenomena can be observed both within constituents and within the sentence: the examples given above ("grow", "eat") concern the interaction between a subject noun phrase and the main verbal predicate of the sentence. Similarly, we observe selection phenomena between verbs and their subcategorized complements, e.g. objects, prepositional objects, etc., but also with adjuncts, or within other constituents than VPs, for example in adjective phrases (noun + attributive adjective).

Collocations can be classified, at least for languages like English, the Germanic, Romance and Slavic languages, according to the category of their elements, into noun–verb, noun–adjective, noun–noun collocations, as well as verb–adverb and adjective–adverb. Noun–verb collocations can be further subclassified according to the grammatical function of the noun phrase contributing the noun part of the collocation: subject–verb–, verb–complement–, verb–adjunct–collocations.[4] Following Hausmann (1979), Hausmann (1985) and Hausmann (1989), we have classified a few examples in the illustration in Figure 2.

NOUN + adjective	*confirmed bachelor*	*eingefleischter Junggeselle*	*célibataire endurci*
NOUN + verb (Subj)	*his anger falls*	*Zorn verraucht*	*la colère s'apaise*
NOUN + verb (Obj)	*to withdraw money*	*Geld abheben*	*retirer de l'argent*
VERB + adverb	*it is raining heavily*	*es regnet in Strömen*	*il pleut à verse*
ADJ + adverb	*seriously injured*	*schwer verletzt*	*grièvement blessé*
VERB + adverb	*to fail miserably*	*kläglich versagen*	
NOUN + noun	*a gust of anger*	*Wutanfall*	*une bouffée de colère*

Figure 2. Types of collocations in terms of the category of their components, following Hausmann (1989)

This notion of collocation does not assume that all collocations make up phrases: n+adj–collocations may do so, if the adjective is used attributively, as in EN *heavy rain*, EN *unquenchable thirst*, DE *starker Raucher*, FR *regrets amers*, FR *remords tardifs*, etc. However, we still want to consider the combination of EN *unquenchable* and *thirst* as collocational, when the adjective is used predicatively (*His thirst...was...unquenchable.*). This implies, among others, that computational tools which would just look for combinations of adjacent lexemes,[5] would not retrieve all combinations which fall under the syntactic definition given above.

The noun which participates in an n+v collocation can also be located in an *adjunct* (cf. DE *es regnet* ***in Strömen***, etc.); such cases are difficult to treat in a strictly valency–based model or in a formal account which makes use of subcategorization information only. To our knowledge, not much work has so far been done on "(lexically) typical adjuncts".

As observed, combinatory phenomena are often orthogonal with phrase structural or valency–based grammatical rules (in the widest sense). This property is problematic for example for lexical choice in natural language generation; in early approaches, the order in which lexemes were selected in a sentence to be generated, was determined by relationships between syntactic heads and modifiers, or nodes of a valency representation and their dependents (rule: "lexical heads first"). This works out for verb–adverb– or adjective–adverb–collocations and for some noun–adjective–collocations as well, but not for noun–verb–collocations (e.g. in the verb–object case: the object noun must be determined first, only then a collocationally adequate

verb can be selected). Researchers in natural language generation were first to discuss problems of collocation: some of the work on lexical choice is aimed at bringing collocational and syntactic constraints together and controlling their interaction in an adequate way (cf. e.g. Nirenburg et al. 1988, etc.).

An additional problem of the interaction between syntactic and collocational description is the recursive nature of collocational properties: the components of a collocation can again be collocational themselves: next to the German collocation *Gültigkeit haben* (n+v), we have *allgemeine Gültigkeit haben*, with *allgemeine Gültigkeit*, a collocation (n+a), as a component. These cases have sometimes been analyzed as different from collocations, but there is no reason for such treatment. However, a formal account, e.g. for machine translation, would have to be able to account for such cases.

2.2.2 Problems of the syntactic description of collocations – the case of support verb constructions

Syntacticians have observed some irregularities in the syntactic behaviour of collocations, in particular of support verb constructions ("Funktionsverbgefüge", "constructions à verbe support"): examples are FR *avoir peur, avoir faim, prendre un bain, poser une question, opérer un choix*, EN *be in a habit, take a bath, pay attention, deliver a speech*, DE *Angst haben, ein Bad nehmen, eine Frage stellen, zur Anwendung kommen.*

Many of the syntactic operations possible with verb phrases are not or only in part possible with support verb constructions; such operations (often used as tests) include passivization, pronominalization, the possibility of taking the nominal part up with an anaphoric pronoun, the possibility of modifying the noun (e.g. with adjectives, genitives, relative clauses, etc.), the choice between different kinds of determiners, etc.

The most "frozen" (or as Cruse (1986:41) says, "bound") collocations are close to typical examples of idioms, insofar as no modifications are possible. Other support verb constructions participate in some, but not all of the processes mentioned above: DE *eine Frage stellen* can be modified or pronominalized, whereas DE *zur Ausführung gelangen* does not allow pronominalization or modification: *Hans hat eine kluge Frage gestellt, Josef hat **sie** beantwortet*; **das Programm gelangt zu einer vollständigen Ausführung; das Programm gelangt zur Ausführung: *sie muß korrekt sein.*

Apparently, pronominalization or pronominal anaphoric reference and the possibility of modification of the predicative noun are somehow related. Similar data to those for German have been observed for Danish (cf. e.g. Dyhr 1980), Dutch (cf. Hinderdael 1980) and French (cf. Gross 1986, Gross/Vivès 1986). In many articles about support verb constructions, just some such facts are described ("anecdotically"), and we are still not aware of a more comprehensive treatment or a formal account. It seems that the

two types of "lexicalized" and "non–lexicalized" support verb constructions observed by Helbig (1984) roughly correspond to cases where the predicative noun is still available as a referent ("non–lexicalized" case) as opposed to referentially "blocked" cases ("lexicalized"): ongoing work by Kuhn (1994) shows that the test used to distinguish these two types are all based on referential (un–)availability.

Similarly, the syntactic (e.g. valency) behaviour of lexical combinations (including both collocations and idioms) has not been described in very much detail so far in dictionaries. We only know of projects for foreign language learners' idiom dictionaries which aim at coming up with a detailed syntactic description.

Noun compounding is often not looked at from the point of view of lexical combinatorics; but a collocational view is most relevant, e.g. for contrastive work on Romance vs. Germanic languages. Mel'chuk's examples of expressions for groups of animals (EN *flock of seagulls, pack of dogs, school of fish*, cf. Fontenelle (1994b) for more examples), but also technical terms like those described and analyzed by Seelbach (1994) (IT *acque di rifiuto* – DE *Abwässer*; IT *stazione di depurazione* – DE *Kläranlage* IT *perdita di sostanze liquide* – DE *Flüssigkeitsverlust*) are cases in point. Knowledge about collocationally adequate combinations of nouns in compounds or noun phrases is most important for translation. Soler/Martí (1994) discuss this problem in detail, giving examples from Spanish–English translation.

2.3 Semantic properties

2.3.1 Combinatory phenomena and compositionality

Syntactic properties do not seem to have much discriminatory power, as far as collocations and idioms, their borderline and the borderline with "normal constructions" are concerned. For tests and criteria of classification, we thus have to rely on (lexical) semantics.

A few general, broad distinctions seem to be commonly accepted: the meaning of idioms is not derivable from the meaning of the lexemes, word forms which make up the idiom: idioms are non–compositional. On the other hand, what has been called "free combination" by Hausmann and others, i.e. the "normal case", is fully compositional: the meaning of EN *to buy a book* is derivable by the usual processes from the meanings of EN *buy* and EN *book*. Collocations are an intermediate case between the two: the meaning of EN *buy somebody's argument* is not fully compositionally derivable from the meanings of *argument* and *buy*. However, the meaning of *argument* is present in (and used in the meaning description of) *buy sb's argument*, it is only *buy* which does not have, in this collocation, the meaning it has in ***buy*** *a book*. This partial compositionality of collocations has led Hausmann to describe collocations as "polar" combinations, consisting of a *base* (the item

which has its full lexical meaning, in our example above: *argument*) and a *collocate*[6] (with modified or "reduced" meaning: *buy*).

Mel'chuk, in a talk about collocations at the 1990 conference of Euralex,[7] has most clearly summarized the differences in compositionality between free, collocational and idiomatic combinations, and we schematize these in Figure 3, following Mel'chuk's presentation.

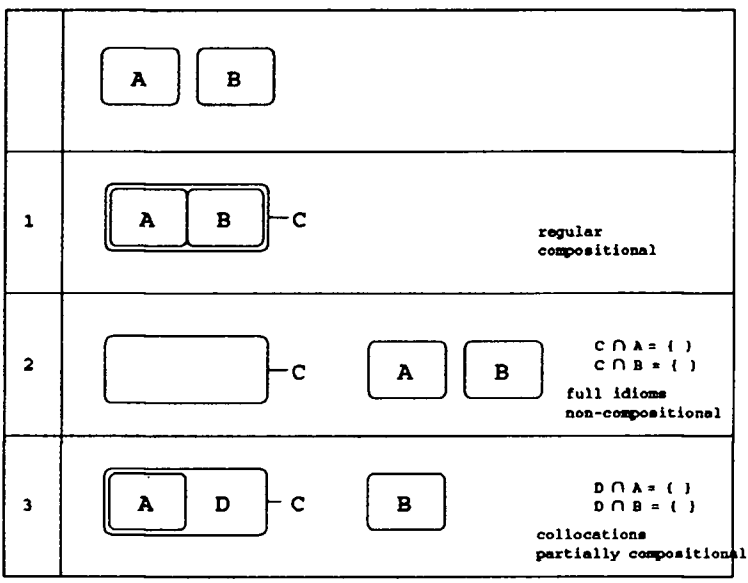

Figure 3. Types of lexical combinations in terms of compositionality (following Mel'chuk)

When constructing semantic representations, we can apply the usual procedures for compositional cases to free combinations; we can assign a single semantic representation to an idiom as a whole. Dobrovol'skij 's examples DE *den Kopf hängen lassen* or *etw. in die Wege leiten* could be described as denoting "resignation" or "startup of an activity" respectively, and we could, depending on the granularity of description we aim at, describe DE *jemand läßt den Kopf hängen* in a similar way as *jemand ist deprimiert* or *jemand ist resigniert*, or as well *jemand leitet etwas in die Wege* similarly to *jemand beginnt etwas, jemand leitet etwas ein.*

A more difficult problem is the following: collocations like DE *zur Anwendung gelangen* ("to be applied") can, as it seems, be represented by the same device as the idioms above: at a certain level of specificity, we can consider such collocations as quasi–synonymous with verbs (DE *anwenden*,

in this case) and use the same representation for DE *angewendet werden* and *zur Anwendung gelangen*, only with an aspectual difference.

But what about the cases where the predicative noun is referentially available, as in the example discussed above, in Section 2.2.2: DE *Hans hat eine **Frage** gestellt. Max hat **sie** beantwortet.* In this case, we need a representation which would preserve an antecedent for the anaphoric pronoun. If the semantic representation is just the same as that of a two–place verbal predicate ("ask": DE *fragen*, in the case of *eine Frage stellen*), no "hook" to serve as an antecedent for the anaphoric pronoun is available. The same way, no reasonable semantic representation of the modifier in DE *er hat eine **kluge** Frage gestellt* would be possible then. If we treat the "referentially available" cases separately, do we need different representations for DE *Verkauf* in *zum Verkauf stehen* (no referent) and *einen Verkauf tätigen* (referent available)? This problem comes up when one tries to give a slightly more formal account of support verb constructions, e.g. in Head Driven Phrase Structure Grammar, HPSG (see e.g. work of Erbach/Krenn 1993), or in any other framework usable in NLP. It also comes up in translation: Thurmair (1990) discusses cases like the translation of EN *to launch (a product)* by DE *(ein Produkt) auf den Markt bringen*; if we use a "compact" semantic representation for the collocation, i.e. one which would be similar or identical with that of the verb, we would be in trouble to translate back from DE *(ein Produkt) auf den **überfüllten** Markt bringen* into English. Other, more detailed semantic representations seem necessary. We have to ask ourselves, then, however, how far we should go in "decomposing" the meaning of collocations, derivatives and of "one–word lexemes".

2.3.2 Towards a semantic classification of collocations? Mel'chuk's lexical functions

The Meaning ⇔ Text–Model (MTM), or: Meaning–Text Theory, developed by Igor A. Mel'chuk and his collaborators includes, among many other things, a semantic classification of lexical combination phenomena. The approach is much broader than just a description of the semantic classes into which collocations can be subdivided: it is a whole theory of language, conceived as a model of how meanings can be realized in language.

It is impossible to give a full and adequate characterization of MTM in the framework of the present article. It is sufficient, here, to recall a few of its most important aspects:

- MTM equally supports analysis and generation of text, but its primary goal is an account of generation, i.e. the problem of how meanings get realized in texts (hence the name "Meaning ⇔ Text–Theory"). Consequently, the description of paraphrasing, of quasi–synonymy, of

the shading of meaning, depending on communicative and text–structural phenomena, etc. are important to MTM researchers;
- MTM is a modular and stratified approach. It distinguishes several strata, roughly corresponding to the traditional levels of description and representation (semantic, "deep" and surface syntactic, morphological and phonological);
- MTM syntax is dependency–based; descriptions of verbs in the dictionary include an inventory of the relevant actants and of their syntactic realization (e.g. as phrase structural constructs); this is a basis for the definition of a syntax–semantics interface, and it also allows to link the description of collocations to this interface.[8]

MTM describes collocations by means of "lexical functions". These can be seen as relations between one or more words or word combinations on the one hand and a partial semantic description on the other. The partial semantic description consists of a "keyword" and an abstract semantic operator applied to this keyword; the different kinds of operators are the different types of collocations.[9]

The number of lexical functions is limited to around 60; they can be combined (see e.g. Ramos/Tutin 1992, work by Mel'chuk, etc.). Out of these 60 lexical functions, about a dozen play an important role to describe collocations of indoeuropean languages.[10]

The table in Figure 4 contains a few examples of lexical functions, along with the name of the LFs and our very rough description of the "meaning" expressed by each of the operators.

"Meaning"	lexical function	examples (French, English)
Intensifier	MAGN	*bruit infernal, interdire absolument*
"Quantity selectors"	MULT	*un essaim d'abeilles*
	SING	*un grain de riz, a cake of soap*
"Evaluator":	VER	*sharp knife, merited praise*
semantically (almost)	EPIT	*océan immense*
empty stylistic	GENER	*un sentiment de joie*
figures	FIGUR	*un rideau de fumée*
points in a process	GERM	*seed of hope*
	CULM	*paroxysme de joie*
semantically (almost)	$OPER_1$	*porter plainte, pousser un cri, mener une lutte*
empty support verbs	$OPER_2$	*sth. forms an offer, sth. constitutes an offer*

Figure 4. Examples of lexical functions

The Meaning ⇔ Text–Model is not only a framework for the description and semantic classification of collocations: Mel'chuk and his collaborators have also worked out proposals for very detailed dictionaries, the *Explanatory and Combinatory Dictionaries* (ECD); these proposals[11] have

been most stimulating for both lexicography (cf. the dictionary of Cohen (1986), the three volumes of French ECDs, as well as ECD fragments of Russian and studies towards ECDs of English (Steele 1988), and German) and Natural Language Processing (cf. work by Nirenburg et al. 1988, Heylen/Maxwell 1994, Heid/Raab 1989). Still some deficits have been identified, such as the fact that the level of granularity of the semantic description of lexical functions may not be fully sufficient for semantics–based NLP (cf. Heylen/Maxwell 1994) or the lack of a "grammar" for combining lexical functions among; this latter gap has been filled by Ramos/Tutin (1992).

If one compares the MTM approach to collocations with the work of Hausmann and other lexicographers, as Cop (1990) and Heid (1992) have done, quite some overlap is found, despite differences in terminology. The table in Figure 5 comparatively summarizes the relevant terminology used in Mel'chuk's and Hausmann's work.

Compared	Who?	components/properties	
Terminlogy	H. M.	Base Keyword	Collocate Value of LF
Semantic properties	H. M.	autonomous compositionally describable	dependent, non-autonomous not fully compositionally
Implication for treatment of collocations	H. M.	collocations must be learned separately collocations must be stored explicitly in the ECD	

Figure 5. Comparing terminology of MTM and Hausmann

Collocation–related research topics in MTM include the actual integration of a collocational component into implementations, as well as work on the relationship between semantics and collocation (see Section 2.3.3 below).

2.3.3 Correlating semantic classes and collocational behaviour

It has been stated that collocations and collocational lexical choice are completely lexically determined (cf. Mel'chuk/Polguère 1987) and thus need to be memorized, by foreign language learners (cf. Hausmann 1984) or in

dictionaries, be it for human use or for NLP. On the other hand, some research has been going on, over the past few years, about correlations between semantic classifications, lexical fields, etc. and collocational behaviour. Heid/Raab (1989) have observed that the French nouns denoting personal attitudes which are described in the first volume of the ECD (cf. Mel'chuk et al. 1984) select similar collocates, for certain lexical functions: for a dozen of semantically related nouns,[12] a parallel behaviour in collocate selection for the lexical functions OPER_1, CAUS OPER, INCEP FUNC, INCEP OPER, FIN OPER, ... was observed.

In the field of lexical acquisition, it has been tried to constitute lexical semantic classes (or domain classes?) by considering collocational behaviour: the assumption is that bases having the same collocates belong to the same field. Pustejovsky et al. (1993) have used this assumption in terminology–related corpus exploration.

Much more material is now analyzed in studies by Meyer/Mackintosh (1994) and in particular by Mel'chuk/Wanner (1994). While the first is on sublanguage collocations, the second deals with general language, coming back to the field of emotion nouns, subdivided, for that purpose, into (in part overlapping) subsets, according to inherent properties of emotions, as they are described in psychology: positive vs. negative emotions, moderate vs. intense, temporary vs. permanent, etc. For each such subset, the collocate selection behaviour of a few prototypical German base nouns and selected lexical functions is analyzed.

The results are of two types: on the one hand indeed, a number of collocations appear with most or all of the elements of the field or of a given subset; on the other hand, a non–negligible amount of exceptions is noted as well. Mel'chuk/Wanner (1994) take this result as a starting–point for a proposal for the reorganization of the ECD entries for emotion nouns. The proposal is to introduce a common "public" entry for the whole class of nouns which would stipulate the values of certain lexical functions, either for all of the class members, or in function of the presence of one or more of the subclass–defining criteria. The results do not immediately lead to a hierarchy; the domain model used is not hierarchical neither. What comes out are rather implications between the presence of certain semantic properties and the collocation behaviour.

Meyer/Mackintosh (1994) observe that, for the sublanguage of technical documentation of CD–ROM devices, a few collocational generalizations are possible, which can be modeled in an inheritance hierarchy. Maybe in part the differences between Mel'chuk/Wanner's and Meyer/Mackintosh's results have to do with the fact that terminological domains, especially when denoting concrete objects, can more easily be modeled themselves in taxonomies than domains of abstract notions, as used in general language. The result is interesting in the light of Martin's notion of 'conceptual collocation': Martin (1992) observes a correlation between the semantic and conceptual description of items of a (technical) domain and the collocational

behaviour. He observes a subtype of n+adj– and n+n–collocations which just denotes (specialized) subtypes of the objects denoted by the base noun. Similarly, he points out that n+v–collocations denote what one can typically do with (or to) the object denoted by the base noun.[13]

We have made a few experiments on this problem ourselves, using Cohen's description of collocations of the sublanguage of the stock market as a starting point (Cohen 1986).[14]

We have looked up the entries for nouns which share certain collocational properties. One type of question we asked was to know which subsets of nouns share one or more collocate expressing the INCREASE or DECREASE of the process denoted by the base noun, both with subject– and object– taking verbs. One such group consists of *<hausse, baisse, mouvement, progression, recul, repli, reprise>*: all these nouns share the collocates *<(s')amplifier, (s')accélérer, (s')accentuer>* for the INCREASE (pronominal reflexive verbs taking the nouns as subjects, the other verbs as objects), *<(se) ralentir>* (verb taking the nouns as subjects) and *<limiter qc., freiner qc.>*, to express the DECREASE. It is not hard to identify properties of this subset of nouns: they all denote changes in the economic evolution or just an economic evolution itself. The subset is quite homogeneous. So is another subset: *<action, change, indice, titre, valeur mobilière>* which take *<monter, augmenter>* to express an INCREASE and *<baisser>* to express a DECREASE (all verbs with nouns as subject). These denote shares or indices. The illustration in Figure 6 shows a few more such clusters.

cluster/collocates → ↓	INCREASE noun=subj	noun=obj	DECREASE noun=subj	noun=obj
achat, concurrence, déficit, dépense, emprunt, épargne, excédent	s'accroître, augmenter	accroître augmenter	diminuer	restreindre
action, change, indice, titre, valeur, mobilière	monter, augmenter	baisser		
économie, balance des paiements	s'améliorer	–	s'affaiblir, se dégrader se déteriorer	affaiblir affaiblir
activité, bénéficier, chômage, coût, commande, consommation, cours, demande, dividende, exportation, importation, investissement, marché, marge bénéficiaire, masse monétaire, offre, pouvoir d'achat, production, productivité, profit, taux, vente	s'accroître, augmenter,	accroître	baisser	freiner, réduire
charges, concurrence, déficit dépenses, déséquilibre, dette, écart, échanges, emprunt, épargne, excédent, impôt, liquidités, perte, plus-value, rendement, réserves, ressources, volume des échanges, volume des transactions,	s'accroître, augmenter	accroître	diminuer	diminuer

Figure 6: Nouns sharing verbal collocates, retrieved from Cohen (1986)

The result of this exploration shows, among other things, the following: some collocate verbs are "passe-partout", like <*s'accroître, augmenter*> denoting the INCREASE; other verbs are selective, e.g. between nouns denoting economic events perceived as "action-like" (and thus collocating with <*freiner,réduire*>, for example) and events perceived as situations or states (subclass selected by e.g. <*diminuer*>).

Outpreliminary exploration seems to lead to results compatible with Gaston Gross's notion of "object classes" (cf. Seelbach (1994) for examples and discussion) and to confirm the assumptions about the relationship between semantics and conceptual collocation. There seems to be a gradual difference between 'conceptual collocation', predominant in terminology, and plain 'lexical' collocation, predominant in general language; the fragment covered by Cohen (1986) is likely to be a borderline case.

Another question we can ask concerns the "selectivity" of less common collocates. We have looked at *bondir* and *s'effondrer*, for INCREASE and DECREASE, respectively (cf. the tables in Figure 7 which display both overlaps and differences in the subsets of nouns sharing these collocates), and at the subset of nouns taking both *s'améliorer* and *se dégrader* for INCREASE and DECREASE: these are *conjoncture, économie, balance des paiements, équilibre, pouvoir d'achat, tendance*, all expressing states or economic situations. The noun FR *emploi* is also a member of this group. *Emploi* is defined by Cohen (1986) as follows: "*Somme du travail humain effectivement employé et rémunéré dans un système économique*" and thus well belongs to the group of economic situations.

Base	Verb + SUBJ DECREASE = s'effondrer
action	+
cours	+
indice	+
marché$_2$	+
monnaie	+
prix	+
titre	+
valeur immobilière	+

Base	verb + SUBJ INCREASE = bondir
cours	+
exportation	+
importation	+
indice	+
monnaie	+
prix	+

Figure 7. Nouns sharing *s'effondrer* or *bondir* as collocates: nouns in boxes are members in both subsets.

Which conclusions can we draw from this small experiment? We will most likely not be able to obtain just one hierarchy of semantic subclasses by clustering nouns of a given domain according to shared collocates. But certain groups can be identified nevertheless, and the sharing of collocations

in such small groups of nouns is significant and most likely can be related with properties relevant for the semantic or conceptual description of the group of nouns.

Terminologists and lexicographers might usefully explore in more detail the relationship between conceptual or semantic description and collocational behaviour. As Martin (1992) states, results of a collocational analysis furnish input for definition construction and vice versa, definitions (in terms of frames, for example) can be used as a background for collocational expectation patterns.

With the availability of corpus processing tools, such analyses become less expensive. Cohen did not explicitely group the nouns treated in her dictionary, although this would be possible, as our experiments show and as the work of Mel'chuk/Wanner (1994) and Meyer/Mackintosh (1994) suggest. Such structuring would be helpful for pedagogical purposes. Knowles/Roe (1994) deal with the pedagogical use of collocational material extracted from texts of specialized language; some of the tools described by Grefenstette (1994) are helpful for technically doing the job. Tools, methods, applications and descriptive work come together at this point: affaire à suivre.

2.4 Pragmatic properties

The pragmatic description of collocations involves the notion of collocations as conventionalized expressions. General language collocations are "the normal way" of expressing a given meaning (cf. *sich die Zähne putzen/*bürsten* vs. *se brosser/*nettoyer les dents*). Hausmann calls collocations "semi–finished products" of language ("Halbfertigprodukte der Rede"). This is why collocationally correct texts are perceived as "fluent", whereas texts with wrong collocates or with compositional expressions where collocational alternatives would exist, are perceived as unnatural; this property of collocations in turn motivates much of the pedagogical interest they attract.

In addition, individual collocations can pertain to diasystematic language varieties, the same way as one–word lexemes can (cf. Swiss German *einen Entscheid fällen* for German *eine Entscheidung treffen*; "East German" *eine Bestellung auslösen* vs. *eine Bestellung aufgeben*).

3. Lexicographic treatment of combinatory phenomena – access to collocations

We have so far mentioned a few problems of the description of collocations. In addition, the properties of collocations lead to a number of particular problems concerning the presentation in dictionaries of collocational descriptions. Here, we capitalize on the organization of lexical entries and the access to collocational information in dictionaries.

Although, from a semantic point of view, it would probably be a good solution to have individual lexical entries for collocations and idioms (and to make them accessible as a whole), this is not practical within semasiological dictionaries. This is, however, what happens in onomasiological dictionaries, such as the *Longman Language Activator* (LLA) or the dictionary of idioms planned by Dobrovol'skij (1994).

3.1 The organization of collocation and idiom dictionaries

Lexicographers have much discussed the access to idioms and collocations in monolingual and bilingual dictionaries; in particular the question where to alphabetize multiword expressions: this problem must be solved in different ways, depending on the distinctions between monolingual and bilingual dictionary and between encoding (text production) and decoding (text understanding) use of the dictionary. Production dictionaries will favour the access to collocational information via the base, whereas in a decoding dictionary we can not be sure that the reader of a text is able to figure out whether or not a wordform belongs to a collocation, and, access via both, bases and collocates, or via the collocation as a whole would be ideal.[15]

This is easier to realize online; an experiment of this type has been made in a lexical and terminological database designed to hold single word items as well as collocations, which has been designed by Heid/Freibott (1991).

The following problems of access to combinatory information in dictionaries have been discussed in the literature.

For idiomatic expressions, the problem is particularly hard, since usually none of the word forms which make up the idiom is a clear candidate, on semantic grounds, to serve as an entry word.

For collocations, Hausmann (1988) has suggested to sort them under the bases. This is what happens consistently in Ilgenfritz et al. (1989) (cf. the entry for *respect* in Figure 9). This sorting procedure is in line with the tradition of stylistic dictionaries, such as Lacroix (1956) and others. An example of an entry from Lacroix (1956) is reproduced in Figure 8. It lists the verbal and adjectival collocates of the entry word, sorting them in part according to their subcategorization properties. The *BBI combinatory dictionary of English* (Benson et al. 1986) also organizes its macrostructure by the bases treated, listing the collocates in the body of the entry.

Respect. Éprouver, ressentir, montrer, marquer, témoigner, manifester, devoir, porter, professer, affecter, feindre du respect. Inspirer, provoquer, commander, forcer le respect. Manquer de respect. Adresser ses respects. Être entouré d'un certain respect. Rappeler au. — QUAL.: profond, filial, sincère, craintif, général, unanime, universel.

Figure 8. The entry s.v. *respect* in the collocation dictionary by Lacroix (1956)

respect m *Respekt, Achtung*
avoir, ressentir du ~ envers, pour, à l'égard de qn *j-m Achtung entgegenbringen:* Nous ressentons du ~ envers Monsieur votre père. / **devoir** le ~ à qn *j-m Respekt schulden:* Nous devons le ~ à nos professeurs. / **forcer** le ~ de qn *j-n Achtung abnötigen:* Son comportement a forcé mon ~. / **imposer, inspirer, commander** le ~ *Achtung einflössen:* Cette personne, bien qu'elle soit très petite, inspire le ~. / **manquer** de ~ (envers qn) *es an der notwendigen Achtung fehlen lassen (gegenüber j-m:)* Je trouve qu'il manque de ~ envers ses parents. / **témoigner, montrer** du ~ à, envers, pour, à l'égard de qn *j-m Respekt erweisen:* Les enfants d'aujourd'hui ne témoignent plus tellement de ~ aux persones âgées.

Figure 9. The entry s.v. *respect* in Ilgenfritz et al. (1989)

A quite detailed syntactic account of collocations similar to our proposals in Figure 2 is given in Lainé (1993): this dictionary (specialized vocabulary of CAD/CAM French/English) distinguishes subject–verb–, verb–object–, and noun–adjective–collocations, as well as collocational noun phrases involving PPs (compound nouns). Below, we reproduce an example of an entry. It consists of two columns, one of which contains the syntactic classification used in the dictionary, the other the relevant lexical combinations.

ordonnancement	**scheduling**
~ V.	*~ connaître les ordres lancés*
V. ~	*choisir ~, définir ~, essayer ~, [règles] gouverner ~*
~ Adj.	*~ assisté par ordinateur, ~ dynamique, ~ informatisé, ~ multiconvergent, ~ optimal*
~ (Prép)(Art)N	*~ à buts multiples, ~ par dates croissantes, ~ par valeurs croissantes des marges libres*
N(Prép)(Art) ~	*l'art de l'~, coefficient d'~, fonction ~, méthode d'~ (de production), rebouclage sur l'~, technique d'~*

Figure 10. The entry s.v. *ordonnancement* in Lainé (1993)

3.2 Finding collocations in general dictionaries

The dictionaries which we have considered in Section 3.1 are all specialized collocation dictionaries. General dictionaries do contain collocations, but sometimes have a much less clear policy for the lemmatization of collocations. Most monolingual definition dictionaries do not have a separate item type (Wiegand's terminology) to indicate collocations. A case in point is the *Oxford Advanced Learner's Dictionary* (OALD), which, in its third (electronic) edition, distributes collocational information or examples of collocations over the items giving definitions or

(glossed) examples, as well as subentries. Similarly, *Cobuild* has collocations in its definienda, as well as in its examples.

Among bilingual dictionaries, the Collins/Robert English/French dictionaries and the Collins/Klett English/German ones are a remarkable exception: they have particular devices to denote n+v–collocations, distinguishing even whether the noun is the subject or a complement of the verb. These dictionaries have been collocationally explored and described in detail by Fontenelle (1992a) etc.: on top of their well–structured representation of collocations, they are a remarkably rich source for this type of lexical information. Another particular device for the treatment of collocations has been used in the Van Dale bilingual dictionaries. They follow the idea of a categorial description of the component parts of collocations and indicate, for example, verbal collocates of a noun in a special part of the entry, using a numeric code to point to the category of the combination partner.[16] A sample entry from the FR –> NL dictionary is reproduced in Figure 11.

respect <f> <m.> **0.1** ***erbieed*** ⇒ ***(hoog)achting, ontzag, respect*** **0.2** ***eerbieding*** ⇒ ***naveling*** **0.3** <mv.> ***betuigingen van hoogachting*** ◇ **2.1** ~ humain ***vrees voor wat men ervan denken, zeggen zal*** **2.3** mes respects à votre femme ***de groeten aan uw vrouw;*** <mil.> mes respects <begroetingsformule v. onergeschikte tgov. officier> **3.1** avoir du ~ pour qn. ***achting, respect voor iem. hebben;*** commander, imposer, inspirer le ~ ***ontzag inboezemen, respect afdwingen;*** manquer de ~ à, envers qn. ***zich tegenover iem. onbehoorlijk, niet correct gedragen;*** manquer de ~ à une femme ***zich vrijpostig gedragen tegenover een vrouw;*** montrer, témoigner du ~ à, envers, pour qn. ***iem. achting betonen, betuigen;*** garder, tenir qn. en ~ ***iem. in bedwang houden, iem. onder schot houden*** **3.3** présenter ses respects à an. ***iem. de groeten doen*** **4.1** ~ de soi ***zelfrespect*** **6.1** **sauf** le ~ que vous dois, **sauf** votre ~ ***met uw verlof, met uw welnemen, met alle respect.***

Figure 11. The entry s.v. *respect* in the Van Dale FR/NL dictionary

3.3 The ECDs: access via semantic criteria

The above dictionaries, specialized and general, monolingual and bilingual, use syntactic criteria for the organization of collocational information. The only dictionaries we are aware of to base their organization on semantic criteria as well, are the *Explanatory and Combinatory dictionaries* which have been published by Mel'chuk and his research group, such as Mel'chuk et al. (1984), etc. The access to collocations is via the base entry and the lexical functions applicable to the base entry (see Section 2.3.2 and note 9). A small part of the entry s.v. *respect* is given in Figure 12.[17]

$Oper_1$	avoir, éprouver [ART ~] *[Toute la population a <éprouve> un profond respect pour cet artiste émérite]*
continuellement $Oper_1$	vivre [dans le ~] *[Cette famille vit dans le respect de ses ancêtres]*
$ContOper_1$	garder [ART ~] *[Malgré les propos diffamatoires des journalistes envers ce député, ses proches collaborateurs ont gardé un profond respect pour lui]*
$FinOper_1$	perdre [ART ~ / tout ~] *[Les dirigeants ont perdu tout respect pour ces artistes]*
$Caus_{(3)}Oper_1$	inciter [N à ART ~] *[Les parents les incitent au respect des valeurs morales; L'honnête de Louise incite Paul et Jean au respect de cette femme]*
$nonOper_1$	ignorer [tout ~ <l'idée même de ~, toute forme de ~ >] *[Jean ignore tout respect pour ses parents]*
$Oper_2$	jouir [de ART ~], avoir [le ~] *[Pierre jouit du respect de ses subordonnés]*
$ContOper_2$	conserver [le ~] *[Malgré les propos diffamatoires des journalistes envers ce député, ce dernier a conservé le respect de ses proches collaborateurs]*
$FinFunc_0$	disparaître *[Le respect du public pour ce ministre a disparu]*
$Caus_2Func_0$	se mériter [ART ~] *[Par son travail consciencieux, il se mérita le respect de ses collègues*

Figure 12. A fragment of the collocation part of the entry s.v. *respect* in the ECD

An application of the ECD description technique is found in Cohen's dictionary (Cohen 1986) of collocations of the sublanguage of economy (stock market and conjuncture).[18] Instead of using lexical functions, she uses paraphrases of a relevant subset of these; given that many of the items serving as entry words in Cohen's dictionary denote processes, the dictionary indicates phases of the processes, like the start, increase, decrease and end. We reproduce in Figure 13 a part of the entry for FR *emprunt* as an example.

emprunt	nouns	(subj. of) verbs	(obj. of) verbs	adjectives
START	émission lancement		émettre lancer	
INCREASE	accroissement augmentation	s'accroître augmenter monter	accroître augmenter	considérable élevé gros
UNDETERMINED				
DECREASE	baisse diminution réduction	baisser diminuer	réduire restreindre	petit
END			clore, liquider rembourser restituer	

Figure 13. The entry s.v. *emprunt* in the dictionary by Cohen (1986)

The two–dimensional presentation of the material in Cohen (1986) supports access via different ways: the user starts with a base lemma and then can either look up collocations in terms of the phases of the process denoted by the noun, selecting thereafter the adequate grammatical realization, or, alternatively, by jointly using both, semantic and grammatical properties.

3.4 Summary, new proposals

The table in Figure 14 contains an exemplary summary of the types of information we can find in dictionaries and of the ways how this information can be accessed.

Information given in dictionaries	Example cited	Access via ...	Example cited
a collocation is used: it is attested	any	definitions, examples, etc.	any
collocation related with a given reading of the base (explicitely)	Van Dale bilinguals	base + reading (number)	Van Dale bilinguals
category of base and collocate	Van Dale, [Ilgenfritz e.a. 1989] [Lainé 1993]	base + reading + cat. code	Van Dale
for N-V collocations: gramm. function of N	Robert/Collins [Cohen 1986]	base + r., + position (markup)	Robert/Collins
semantic classification of collocations	ECDs, [Cohen 1986]	base + r., lexical f.	ECDs, [Cohen 1986]

Figure 14. Main Features of collocation treatment and access in dictionaries

New proposals or simply new practical solutions come up in dictionaries which consistently follow an onomasiological view. This is discussed, with a view to the plans for a Russian/English idiom dictionary in the paper by Dobrovol'skij (1994). The author uses a set of local (or: partial) conceptual hierarchies, inspired by prototype theory, to organize the "backbone" of the dictionary, and he then "links" the idioms described in the articles to the nodes of these hierarchy trees. A similar approach is followed in the *Longman Language Activator* an onomasiological dictionary for language production. The LLA has collocations, idioms and "normal single word lexemes" as entries, all related by a semantic superstructure spanning up small hierarchies, for about 1,000 "topics". Collocations and idioms are treated here on a par with other lexemes. Access is by the meaning of the multiword item as a whole.

Other proposals for dictionary structure are made in Oubine's (Oubine (1994) bilingual Russian–English dictionary of lexical intensifiers (cf. Mel'chuk's lexical function MAGN). Given that a bilingual dictionary is aimed at, and that not for all collocations full equivalence can be stated, the author opts for keeping the two languages separate, pointing from bases of one language to their equivalents in the other language; the base entries contain alphabetical lists of intensifiers, each with examples and, optionally, usage notes. Access to collocations is normally given via the bases, a "reverse" part can be accessed by the intensifiers themselves, leading to an index of bases modifiable by a given intensifier. Another Russian/English collocational dictionary is described in Benson (1994).

Collocations seem to require a multidimensional description (syntactic, semantic, pragmatic, relation with the domain model, etc.). Representing this information and making it accessible, also indifferent 'ad hoc' combinations, makes a flexible dictionary structure necessary, as it is best achieved with computational tools. Defining the structure of a computational collocation dictionary in a way going beyond the simple encoding in Heid/Freibott (1991) is an interesting task.[19]

4. Acquisition and application of collocational information

4.1 Acquiring collocational information from text

Much of the linguistic and lexicographic discussion about collocations has long been based on a few examples, mostly made up by linguists for the purpose of exemplification. Researchers thus felt that there is a need for lists of collocations collected from textual material. Others, like Fontenelle (1992a), Fontenelle (1992b) have developed and used computational tools to identify collocation candidates in dictionaries and to extract collocation lists from machine readable dictionaries.

On the other hand, practical lexicographers themselves need tools for corpus analysis which would give access to collocation candidates: collocational description in dictionaries is an area where still improvements are possible and necessary, and, on the other hand, the availability of collocation lists extracted from texts is a great advantage for the semantic description of lexical items.

Some of the tools for extracting collocations from texts are based on statistical methods. A few statistical measures of similarity have been used to identify how often words appear together and how similar the contexts are where these words appear. The measures most frequently used are "mutual information", "t-score" and "z-score". There are as well other similarity measures which have been applied in tools for corpus exploration.[20]

We can not, in the framework of this article, discuss the different statistical tools in all detail, and we will thus restrict ourselves to an informal account of the workings of the most simple statistical tools; a discussion of the choices which lexicographers and computational linguists have to make when applying statistical tools with a view to the retrieval of collocation candidates from text.

On that basis, we can identify a few tasks for both research and tool development: essentially the impression is that the combination of both statistical measures and linguistic information (e.g. from pre- and post-analysis steps) is a successful practical way forward.

4.1.1 Simple statistical measures

The most "prominent" statistical measures used, implementations of which are available to many lexicographers, are the "mutual information" index and the "t-score" test.[21]

4.1.1.1 Mutual information

The mutual information index (MI, for short) is used to measure the association between two words; for a given corpus, we can count how often a given word occurs in that corpus (frequency). We can do such statistics for all word forms in a corpus (and use for example the information on very frequent or very rare words to decide whether they should go into a dictionary built for that corpus). By dividing the frequency of a word form by the number of word forms in the corpus, we get the lexical probability of the word form (how often, in relation to the overall number of word forms in the corpus, does it appear?).

When measuring mutual information, we do not only observe the probability of single word forms, but also the probability of combinations of two words ("bigrams"). This leads to three probability values which can be compared:

- the probability of the first word form, w_1: $P(w_1)$;
- the probability of the second word form, w_2: $P(w_2)$;
- the probability of a pair (w_1, w_2) built up of the two word forms: (P (w_1, w_2)).

These three values are compared: the probability that w_1 and w_2 cooccur (e.g. next to each other) is divided by the product of the individual probabilities of w_1 and w_2 each.[22] When computing MI, the "window" (or: "span") within which the word forms have to appear to be taken as "cooccurring" can be defined by the user. We may look just at adjacent word forms or at word forms which are at up to 5, 3, 4, ... word forms distance one of each other.

It is quite evident what the comparison will tell us: the MI value is high, when most of the occurrences of a given item are in fact *co*occurrences with the second item selected: if we compare MI for a set of pairs of, say, noun and adjective, by keeping the noun constant, the adjectives which most typically cooccur with our noun will have the highest MI index. "Typically" means: if the adjective is used next to a noun at all, it is very likely that it is the noun for which the MI is high. The famous *célibataire endurci* should be easy to detect in texts this way.

MI is dependent on the frequency of word forms. If we calculate MI for a list of adjectives, collocating with a noun, we do not see, however, in the MI values, whether the adjectives are frequent or not. With MI, we have to face the "sparse data problem": given that MI is a similarity (or typicality) measure, it will yield high typicality values in cases where an item is rare but cooccurs (by chance or by rule) with another one, in each of its rare occurrences in the text. Rare items may thus be ranked much higher than one would intuitively like them to be.

4.1.1.2 T-score

This problem of "high ranking" of rare forms can be remedied by use of the t-test. The t-test operates on pairs of words. It finds those additional words which are more likely to cooccur with one of the two words from the pair than with the other. The results of the t-test come as positive and negative values. The highest and the lowest values are significant: they indicate strong association with one or the other word. T-score is also not reliable for low frequencies.

It tends to indicate the frequent words that cooccur with the target words, and it allows to separate near synonyms by showing frequent combination partners. Church et al. (1991) have applied it to discriminate EN *strong* and *powerful* by finding out nouns which frequently cooccur with one of these adjectives.

4.1.2 Choices in applying statistical measures for identifying collocations

The statistical measures described above, as well as modifications thereof, are often used to identify collocation candidates in texts; such use depends of course on a number of assumptions and choices.

- *The distance of items compared.* MI and t-score can be calculated on immediately adjacent items or on items occurring within a certain "window" or "span" (cf. the terminology of Sinclair 1991, Clear 1994).
- *The impact of text structure.* The measures can be calculated on bigrams within or across sentence boundaries. Usually, we would assume that limiting ourselves to occurrences within one sentence would lead to more relevant results than ignoring sentence boundaries.[23]
- *The impact of lemmatization and categorial information.* The discussion above, in Section 2.2 led to the assumption we can describe collocations in terms of category combinations (n+v, n+n, n+adj, adj+adj, v+adv), in part even of partial syntactic structures, such as "verb+object", "verb+subject", "noun+attributive adjective", etc.

  Most of the work done so far in using MI and t-score was performed on English material. The impact of word from variation there is not as important as with inflecting languages, like German or the Romance languages. For these it seems useful to have an option, in statistical programs to calculate the measures for lemmas rather than word forms. Moreover, it can be useful to carry out the statistical computation only on, say, adjectives appearing next to a noun, or on verbs and their nominal objects, etc.,[24] i.e. to restrict search and statistical computation according to syntactic environments.

There is a range of choices in the application of the statistical tools, and the axis on which these choices can be arranged basically has to do with the amount of linguistic information which is kept track of, either by pre- or postprocessing: The statistical measures may be applied to material selected according to certain linguistic criteria (e.g. by use of concordances), or relevant material is selected according to linguistic criteria from the set of data extracted by statistical processing. Proposals for tool building in view of collocation extraction thus should be staged along with the amount of information available along with the corpus text:

- raw text, possibly with sentence boundaries,
- text with part-of-speech annotations,
- lemmatized and morphosyntactically annotated text,
- text with possibly an identification of noun phrases, verb phrases, etc.

Composite tools bringing linguists and statistics together have been built recently: Smadja's tool, XTRACT, combines statistical measures and some

parsing (Smadja 1993). Similarly, Grefenstette's work is based on part–of–speech tagged and partially syntactically analyzed corpora; statistical similarity and distribution measures are applied after lists of subject nouns and object nouns of verbs have been extracted from the corpus texts. On such material, it becomes possible to statistically validate claims about the interrelationship between semantic closeness and collocational behaviour: two or more items (say: nouns) are taken to belong to the same semantic set or class if they share frequent collocates in the texts; we have discussed this in detail in Section 2.3.3 above. Here, corpus linguistic analysis and work on the description of lexical fragments (e.g. for sublanguages) come together; work of Knowles/Roe (1994) and Grefenstette (1994) deals with these issues. For sublanguage, Pustejovsky et al. (1993) have explored similar techniques of collocational analysis to fill dictionaries and to distinguish senses of lexical items. The latter issue is investigated by Clear (1994) from the point of view of corpus–based lexicography: for a polysemous item, collocate lists are produced by means of a combination of MI and t–score. Then the lexicographers are asked to cluster the collocates intuitively in such away as to get subsets of typical combinations. For a few items from the collocate lists, again typical cooccurrence partners are searched, and thereby the previous intuitive subdivision of the material into "broad sense based" classes done by the lexicographers is either reinforced or weakened.

The statistical measures, as well as tools or tool components developed on their basis are widely available to lexicographers now. The problem, it seems, is similar to that of concordances: for some languages, such as English, the problem is not to get access to machine–readable material, but to filter it in view of finding out what is lexicographically relevant. The statistical tools as well produce lists of material in someway combinatorially related, and the task of the lexicographer is to isolate the relevant combinations. The application of statistical measures to material preselected and/or preanalyzed by means of part–of–speech tagging, lemmatization or possibly partial parsing seems to be more promising than "blind" statistics.

4.2 Combinatory phenomena in translation

When discussing some of the linguistic properties of combinatory phenomena, we have already pointed out the main problems which need to be solved in translation, when collocations are involved; most of these problems are due to the (partial) impredictibility of the choice of collocates. If the contrastive dictionary can easily give equivalents of base lexemes, equivalents of collocates can only be given *within collocations.*

Thus, bilingual dictionaries either should include collocations as entries, or they must have a (consistent) policy for making collocations accessible through their components. Such policies have been discussed in both bilingual lexicography and in research work on machine translation.

Early proposals for the contrastive treatment of collocations, e.g. in machine translation aimed at listing, for a given collocate verb or adjective, possible bases, and for each collocation the translation of the collocate. The disadvantage of this procedure is that the resulting verb entries are very large and unintuitive, because they lead to a large number of "collocational readings" for each collocate. As an example of this problem, consider a few examples of the translation of collocations (cf. Figure 15) with FR *dresser* into German, and, for comparison, a slightly modified version[25] of the entry s.v. *dresser* in the monolingual French *Dictionnaire du Français Contemporain*, DFC (cf. Figure 16). We reproduce a part of the entry of this dictionary because it shows the relationship between the translation problem and the semantic description: it indicates a number of synonyms, most of them for the use of *dresser* as a collocate in a collocation. The translation of *dresser* in German is of course dependent on collocations: DE *aufstellen, ausstellen, aufschlagen,* etc. are not exchangeable.

dresser des baricades	*Barikaden errichten*
dresser un budg	*ein Budget erstellen*
dresser une tente	*ein Zelt aufstellen/aufschlagen*
dresser une contravention	*einen Strafzettel ausstellen*

Figure 15. Collocations with French *dresser* and their translations into German

1. **dresser** [drese] v. tr.
1° *dresser quelque chose*,
le mettre debout, le mettre dans une position verticale ou voisine de la verticale:
Dresser un mât (syn.: PLANTER).
Dresser une échelle contre le mur.
On a dressé une barrière (syn.: ELEVER).
Ils avaient dressé leur tente (syn.: MONTER).
Dresser un monument, une statue (syn.: ERIGER).
Dresser la tête, le buste (syn.: LEVER).
– 2° *Dresser quelque chose*, l'établir, le mettre par écrit:
Dresser un bilan, une liste, un plan, un constat, un procès-verbal.
– 3° *Dresser la table, le couvert*, disposer les couverts pour un repas (syn.: METTRE).

Figure 16. The entry s.v. *dresser* in the DFC

Recent proposals for contrastive collocation dictionaries favour an MTM-like characterization of collocations and use devices similar to lexical functions as an intermediate representation of collocational semantics. Following Mel'chuk/Polguère (1987), for example, Danlos/Samvelian (1992) propose to organize the contrastive collocational dictionary by bases (i.e. have bases as entries) and to list collocations of source and target language according to their semantic subtype, as is the case in the ECD. A sample of the proposed structure is given in Figure 17:

habitude → ***avoir*** (neuter), ***perdre*** (terminative), ***prendre*** (inchoative),
habit → ***be in*** (neuter), ***get out of*** (terminative), ***get into*** (inchoative).

Figure 17. Entries from Danlos/Samvelian (1992)

The information in Figure 17 is used as follows: to translate a collocation (the examples here are support verb constructions), the base is identified, and the collocation is searched in the base entry (e.g. *perdre ... habitude*). The semantic value (i.e. the lexical function) is determined (in this case: "terminative"). Then the English equivalent of the base is searched (*habitude* → *habit*) and from its entry, the collocation expressing the terminative aspect of *habit* (i.e. *get out of a habit*) is retrieved. This is described in detail by Danlos/Samvelian (1992: 22ff).

A problem which is often encountered in translation is what Dorr and others have called "categorial divergence": a meaning expressed by an expression of a given category (say: adjective) is rendered, in the target language, by an expression of another category (e.g. a verb). An example: in an oral conversation, where the goal is to agree on a date for a meeting, I can say EN *I **prefer** next friday.* In German, the most natural translation would be DE *der nächste Freitag ist mir **lieber*** with an adjective *lieber* where English has a verb (*prefer*). These cases of category change are very frequent; sometimes the collocation is the only lexical means available to express a given meaning: FR *se suicider* ↔ EN *to commit suicide.*

A case where "contextual" phenomena, in this case the syntactic environment of the sentence, enforce the choice of a collocation rather than of a verb is illustrated by the following example, where the fact that the source language has a coordination of two passives would make a "rearrangement" of the target sentence necessary, if the collocation had to be avoided.

- *Ces limitations de vitesse sont annoncées à l'avance et puis **rappelées** aux mécaniciens par une signalisation latérale* [from a text of the French railway corporation].

- *Diese Geschwindigkeitsbegrenzungen werden vorweg angekündigt und dann den Lokführern durch Streckensignalen* ***in Erinnerung gerufen.***[26]

In translation dictionaries, only few such cases have been treated; almost no FR → DE dictionary will give DE *in Erinnerung rufen* as a translation equivalent of FR *rappeler*, but we may find FR *rappeler* as an equivalent of DE *in Erinnerung bringen* in a DE → FR dictionary. For Natural Language Processing, both are necessary.

In machine translation research, the problem of categorial divergences has been discussed to some extent. Danlos/Samvelian (1992) give examples of a treatment in a syntactically based framework. A problem which remains open is the semantic description of these cases. Above, in Section 2.3.1 we have discussed the example of the translation of EN *to launch (a product)* by DE *(ein Produkt) auf den Markt bringen* and vice versa of DE *(ein Produkt) auf den Markt bringen* into English (Thurmair 1990); the German collocation allows for modifications which can not (easily) be expressed with the English equivalent; to be able to remedy the situation we need a more detailed semantic description of collocations, one which goes beyond the mere classification of collocations according to lexical functions. Should the ordinary verb (e.g. EN *launch*) receive the same description as the collocation? If not, how to relate the semantic descriptions monolingually, and contrastively?

The application of collocational descriptions in the context oftranslation leads to problems which in part thus allow to formulate monolingual descriptive problems more precisely, and in part pose additional problems, also with respect to lexicographic description and presentation.

Notes

1 The situation is more complicated, however: for example native speakers prefer *le bébé a poussé.*
2 Cf. e.g. Hausmann (1989); Benson (1989).
3 We use examples from English (EN), French (FR) and German (DE) in this paper, and we "prefix" them by language but do not translate them, to avoid an overloading of the presentation.
4 In the table in Figure 2, we have marked the two relevant examples with "(Subj)" and "(Obj)" respectively. We have indicated, in the leftmost column of Figure 2, the element which is determined (the "base", in Hausmann's terminology), in SMALL CAPITALS, the determining element ("collocate") following it. Examples come from EN, DE and FR.
5 Cf. the discussion about "windows size" below, in Section 4.1.2.
6 Hausmann's terms are DE *Basis* and *Kollokator.* See also Benson's summary in English (Benson 1990). In the leftmost column of the illustration in Figure 2, we have typeset BASES in SMALL CAPITALS leaving the collocates in the normal typeface.
7 In Malaga, 1990; the manuscript of this talk has to our knowledge not been published.
8 Much more would have to be said about MTM, but we limit our overview to these points. More details are given by Mel'chuk/Wanner (1994): the annex of their paper contains basic definitions of the MTM apparatus they use in their research.
9 The definition of "lexical functions", e.g. in Mel'chuk et al. (1984), is
"f(X) = Y:

(...) *f* est la fonction lexicale, *X* est son argument (un lexème ou bien une locution), et *Y* est la valeur de la FL *f* pour cet argument, c'est–à–dire l'ensemble des expressions linguistiques qui peuvent exprimer le sens ou le rôle syntaxique donné (noté par *f*) auprès de *X*."

Here is an explanation of the variables used in the formula:

- X (called "keyword (mot clé)" in MTM) is a lexical unit of given language L_1,
- f (called "lexical function (fonction lexicale)" in MTM) is a semantic constant (an abstraction, independent from individual languages or at least generalizable over many languages) which is applied to X, as an operator,
- Y (called "value of a lexical function (valeur d'une fonction lexicale)" in MTM) is a set of actual lexicalizations of L_1.

We can reformulate the definition of the lexical function in a "relational" way: "for a given lexeme *X* of a language L_1, there exists a set of lexemes or combinations of lexemes (e.g. collocations) *Y* of L_1, such that a relation *f* holds between *X* and *Y*, *f* being an abstract semantic operator."

10 This does not mean that they are the only ones whose values come as collocations, but in indoeuropean languages, they serve most often for the description and classification of collocations: Thai and a number of other south–east asian languages have quite regular processes to express 'nomina agentis' through noun–noun–collocations; for the collocational description of these languages, other lexical functions than for German, English or French would play a role.

11 Cf. the discussion of the ECD presentation of collocations, below in Section 3.3 and the sample in the illustration in Figure 12.

12 The items analyzed are *admiration, colère, désespoir, enthousiasme, envie, étonnement, haine, joie, mépris, respect.*

13 The examples from specialized language which we have given above, in the illustration in Figure 1 are mostly of this type: *élution graduée* is a typical subtype of *élution*, in cromatography ("gradient"); the same is true for *ressources renouvelables*, etc.

14 The author should like to thank Regina Steding for preparing a formalized version of the n+v–collocations part of the dictionary in the typed feature structure formalism. Without her work, the exploration we describe here would not have been possible.
A sample entry from Cohen (1986) is reproduced below, in the illustration in Figure 13. The typed feature structures formalism (cf. Emele 1994) allows to formulate underspecified queries more easily than a relational database would do.

15 Cf. Hausmann (1988) for much more details.

16 E.g. **3.x** in the example in Figure 11 indicates a combination with a verb and a second numeric code to relate the collocation with one of the meanings described in the entry; **3.1** points to n+v–collocations of sense **0.1** of *respect*, **3.3** to n+v–collocations of sense **0.3**.

17 We have slightly simplified the items, but preserved most of the layout. The table–like layout is ours.

18 See Cohen (1992) for a discussion of the dictionary and its underlying methodology. See also our discussion above, in Section 2.3.3.

19 The DECIDE project, partly supported by the European Commission, under the MLAP–94 programme of its Directorate General XIII E, Luxembourg, will come up with proposals for such a dictionary structure.

20 Grefenstette (1994) points to a few such measures, in his article.

21 These statistical measures have been described in detail by Church et al. (1991) who have applied them, among others, to identify collocates of "strong" and of "powerful" in English texts.

22 This product would be the "normal probability" of coocurrence of w_1 and w_2. The actual MI index is the $\log_2$:

$$I_{(w1;w2)} = \log_2 \frac{P(w_1, w_2)}{P(w_1) * P(w_2)}$$

See Church et al. (1991) for details.

23 On the other hand, it must be recognized that collocational relations occur also across sentence boundaries; cf. EN: (...) *finally the* ***question*** *of...came up. I would not have dared to* ***ask*** *it.* These cases are most likely not very frequent and leaving them out from the analysis should not be problematic.

24 On the other hand, one may claim that combinatory phenomena need not be a property of lemmas only, and that it may make sense to investigate the combinatory behaviour of individual inflected forms; the same way, it may make sense to include all word classes into the statistical computation, not only the major classes; this latter option is useful for the identification of frequent prepositions, components of phrasal verbs, etc. (what Benson et al. (1986) call grammatical collocation).
25 We have introduced line breaks between the individual collocations or examples and have made sure that each collocation starts with a new line. We have left the typography (fonts) unchanged, because it is relevant for the interpretation of the entries.
26 To test our claim about the syntactic environment, try to translate the sentence without using a collocational equivalent for FR *rappeler.*

References

Benson, M., E. Benson, R. Ilson 1986. *The BBI Combinatory Dictionary of English. A Guide to Word Combinations*, Amsterdam, Philadelphia.

Benson, M. 1989. "The Structure of the Collocational Dictionary", in: *International Journal of Lexicography.* Oxford, Oxford University Press.

Benson, M. 1990. "Collocations and General–purpose Dictionaries" in: *International Journal of Lexicography.* Oxford, Oxford University Press, Volume 3, Number 1, pp. 23–34.

Benson, M. 1994. "Bilingual Collocational Dictionaries." Poster to be presented at Euralex–94.

Church, K.W. et al. 1991. "Using Statistics in Lexical Analysis", in: U. Zernik (ed.) *Lexical Acquisition: Using On–Line Resources to Build a Lexicon.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 115–163.

Clausén, U. and E. Lyly. 1994. "Criteria for identifying and representing idioms in a phraseological dictionary." To appear in *Proceedings Euralex–94.*

Clear, J. 1994. "I can't see the sense in a large corpus", in: Kiefer/Kiss/Pajzs (1994), pp. 33 – 48.

Cohen, B. 1986. *Lexique de cooccurrents; Bourse — conjoncture économique.* Montréal: Linguatech.

Cohen, B. 1992. "Méthodes de repérage et de classement des cooccurrents lexicaux", in: *Terminologie et Traduction*, 2–3, pp. 505–512.

Cop, M. 1990. "The Function of Collocations in Dictionaries", in: Magay/Zigany (1990), pp. 35–46.

Cruse, D.A. 1986. *Lexical Semantics.* Cambridge textbooks in Linguistics, Cambridge.

Danlos, L. and P. Samvelian 1992. "Translation of the predicative element of a sentence: category switching, aspect and diathesis", in: *TMIMT–92, Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation.* Montréal: CWARC, pp. 21–34.

Dobrovol'skij, D. 1994. "Idioms in a semantic network: Towards a new dictionary–type." To appear in *Proceedings Euralex–94.*

Dyhr, M. 1980. "Zur Beschreibung von Funktionsverbgefügen" in: *Kopenhagener Beiträge zur Germanistischen Linguistik. Festschrift für Gunnar Bech.*

Emele, M. 1994. "TFS – The Typed Feature Structure Representation Formalism" in: *Proceedings of the International Workshop on Sharable Natural Language Resources (SNLR).* (in press).

Erbach, G. and B. Krenn 1993. "Idioms and support–verb constructions in HPSG",(Saarbrücken Universität des Saarlandes), ms. [= Computerlinguistik an der Universität des Saarlandes (CLAUS), *Report Nr. 28*].

Fontenelle, Th. 1992a. "Collocation acquisition from a corpus or from a dictionary: a comparison"; in: *Proceedings Euralex–92*, pp. 221–228.

Fontenelle, Th. 1992b. "Co–occurrence knowledge, support verbs and machine–readable dictionaries", in: Kiefer e.a. (1992), pp.137–145.

Fontenelle, Th. 1994a. "Survey of Tools for the extraction of collocations from dictionaries and corpora", ms., (Liège: Université de Liège), Deliverable of the DECIDE MLAP–94 project, 81 pp.

Fontenelle, Th. 1994b. "Discovering significant lexical functions in dictionary entries", paper presented at the International Symposium on Phraseology, University of Leeds, april 1994.

Grefenstette, G. 1993. "Corpus–Derived First, Second and Third–Order Word Affinities". To appear in *Proceedings Euralex–94.*

Gross, G. 1986. "Les nominalisations d'expressions figées" in: *Langue Française.*
Gross, G. and R. Vivès (Eds) "Les constructions nominales et l'élaboration d'un lexique–grammaire" in: *Langue française* 69, 1, pp. 5–27.
Haenelt, K. and L. Wanner (Eds) 1992. *International Workshop on the Meaning–Text Theory,* (Darmstadt: Gesellschaft für Mathematik und Datenverarbeitung mbH), [= Arbeitspapiere der GMD, Band 671].
Hartmann, R.R.K. 1994. "The use of parallel text corpora in the generation of translation equivalents for bilingual lexicography." To appear in: *Proceedings Euralex–94.*
Hausmann, F.J. 1979. "Un dictionnaire des collocations est–il possible?" in: *Travaux de Linguistique et de Littérature* XVII, 1, pp.187–195.
Hausmann, F.J. 1984. "Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen", in: *Praxis des neusprachlichen Unterrichts,* 31–4, pp.395–406.
Hausmann, F.J. 1985. "Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels", in: Bergenholtz/Mugdan (1985), pp. 118–129.
Hausmann, F.J. 1988. "Grundprobleme des zweisprachigen Wörterbuchs" in: Hyldgaard–Jensen, K., A. Zettersten (Eds) *Symposium on Lexicography III, Proceedings,* Tübingen: Niemeyer.
Hausmann, F.J. 1989. "Le dictionnaire de collocations", in: Hausmann et al. (1989),
pp. 1010–1019
Hausmann, F.J., O. Reichmann, H.E. Wiegand, L. Zgusta (Eds) 1989. *Wörterbücher: ein internationales Handbuch zur Lexikographie. Dictionaries. Dictionnaires.* Berlin/New York: de Gruyter, vol. 1.
Heid, U. 1992. "Décrire les collocations – deux approches lexicographiques et leur application dans un outil informatisé" in: *Terminologie et Traduction,* 2–3.
Heid, U. and G. Freibott 1990. "Zur Darstellung von Äquivalenten in einer terminologisch–lexikalischen Datenbank für Übersetzer und technische Autoren" in: Schaeder B., Rieger, B. (Eds) *Lexikon und Lexikographie. Vorträge im Rahmen der Jahrestagung 1990 der Gesellschaft für Linguistische Datenverarbeitung (GLDV)* e.V., Siegen, 26.– 28. März 1990, Hildesheim/ Zürich/ New York, pp. 244–252.
Heid, U. and G. Freibott 1991. "Collocations dans une base de données terminologique et lexicale", in: *Méta* vol. 36, no.1, Montréal.
Heid, U. and S. Raab 1989. "Collocations in Multilingual Generation", in: *Proceedings of the European ACL Conference, Manchester 1989,* Manchester: UMIST.
Helbig, G. (1984. "Probleme der Beschreibung von Funktionsverbgefügen im Deutschen", in: Helbig, G. *Studien zur deutschen Syntax,* Bd.2, Leipzig.
Heylen, D. and K. G. Maxwell 1994. "Lexical Functions and the Translation of Collocations", to appear in: *Proceedings Euralex–94.*
Hinderdael, M. 1980. "Präpositionale Funktionsverbgefüge im Deutschen und im Niederländischen" in: *Studia Germanica Gandensia,* XXI,1980–1981.
Ilgenfritz, P., N. Stephan–Gabinel, G. Schneider 1989. *Langenscheidts Kontextwörterbuch Französisch – Deutsch,* Berlin/ München: Langenscheidt.
Kiefer, F., G. Kiss, J. Pajzs (Eds) 1994., *Papers in Computational Lexicography – COMPLEX '94. Proceedings of the 3rd International Conference on Computational Lexicography, COMPLEX '94, Budapest, Hungary,* Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences.
Knowles, F. and P. Roe 1994. "Facilitating the Corpus–Building Process and Maximising the "Analytical Yield": A LSP–Oriented Case Study", in: Kiefer et al. (Eds).
Krahl, C. 1994 "Aspekte des Kombinationswissens von Verben und Substantiven im Vergleich zu Adjektiven am Beispiel englischer Temperaturlexeme." To appear in: *Proceedings Euralex–94.*
Kuhn, J. 1994. *Die Behandlung von Funktionsverbgefüge in einem HPSG–basierten Übersetzungsansatz,* ms., Stuttgart: IMS–CL, University of Stuttgart, Institut für maschinelle Sprachverarbeitung.
Lacroix, U. 1956. *Les mots et les idées: dictionnaire des termes cadrant avec les idées.* Édition nouvelle, revue et corrigée, Paris: F. Nathan.
Lainé, Cl. 1993. *Combinatory Vocabulary of CAD/CAM in Mechanical Engineering.* Ottawa, Canada.
Magay, T. and J. Zigany 1990. *Proceedings of BudaLEX 1988.* Budapest: Akadémiai Kiadó.
Martin, W. 1992. "Remarks on Collocations in Sublanguage", in: *Terminologie et Traduction,* 2–3, pp. 157–164.

Mel'chuk, I.A., unter Mitarbeit von N. Arbatchewsky–Jumarie, L. Elnitsky, L. Iordanskaja, A. Lessard 1984. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches Lexico–Sémantiques I.* Montréal: Presses Universitaires de Montréal.

Mel'chuk, I.A. and A. Polguère 1987. "A Formal Lexicon in the Meaning–Text Theory (or how to do Lexica with Words)", in: *Computational Linguistics* 13, 3–4, pp. 261–275.

Mel'chuk, I. A. et al. 1992. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches Lexico–Sémantiques III.* Montréal: Presses Universitaires de Montréal.

Mel'chuk, I.A. and L. Wanner 1994. "Towards an Efficient Representation of Restricted Lexical Cooccurrence." To appear in: *Proceedings Euralex–94.*

Meyer, I. and K. Mackintosh 1994. "Phraseological Analysis and Conceptual Analysis: Exploring a Symbiotic Relationship in the Specialized Lexicon." To appear in: *Proceedings of Euralex–94.*

Nirenburg, S. et al. 1988. "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.*

Oubine, I. 1994 "Dictionary of Lexical Intensifiers for Russian and English." Poster to be presented at Euralex–94.

Pustejovsky, J., S. Bergler, P. Anick 1993. "Lexical Semantic Techniques for Corpus Analysis in Computational Linguistics", in: *Computational Linguistics*, Vol. 19, Nr.2, [= Special Issue on Using Large Corpora II].

Ramos, M. Alonso, A. Tutin 1992. "A Classification and Description of the Lexical Functions of the Explanatory Combinatorial Dictionary for the treatment of LF combinations", in: Haenelt and Wanner (1992), pp. 187–196.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford University Press, Oxford.

Seelbach, D. 1994. "Syntagmatic context information for Computer Assisted Translation", ms. Mainz, to appear in 1994.

Smadja, F. 1993. "Retrieving Collocations from Text: Xtract", in: *Computational Linguistics*, Vol. 19, Nr.1, pp. 143–177 [= Special Issue on Using Large Corpora I].

Soler, C. and Martí A. 1994. "Dealing with lexical mismatches." To appear in: *Proceedings Euralex–94.*

Steele, J. (Ed.) 1990. *Meaning–Text Theory, Linguistics, Lexicography, and Implications.* Ottawa: University of Ottawa Press.

Thurmair, G. 1990. "Complex Lexical Transfer in METAL", in: *TMIMT–3*, pp.91–107.

TMIMT–3 1990. Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language, 11–13 June 1990. Austin: University of Texas.