Non-thematic Part

Laura Giacomini

An onomasiological dictionary of collocations: mediostructural properties and search procedures

Abstract
This paper describes a corpus-based case study aimed at modelling an electronic dictionary of Italian collocations addressed to professional translators. The study focuses on a small number of lexical items belonging to the semantic field PAURA (*fear/fright/worry*) and investigates the advantages of an onomasiological arrangement of collocations as well as the organization of dictionary data around a lexical prototype. The intended lexicographic representation, whose function is to support textual production and language skills improvement, rest upon a coherent and solid cross-reference network. The paper pays close attention to this lexicographic mediostructure, which provides the user with exhaustive but also selective information about collocational data.

1. Introduction

This paper discusses the use of an onomasiological approach to the lexicon and a special descriptive metalanguage in the modelling of an electronic dictionary of Italian collocations intended for professional translators[1]. Close attention will be paid to a description of the lexicographic mediostructure and to the various search options that allow specific data to be retrieved and consulted, according to the user's needs. The last part of this paper is entirely dedicated to these focal topics (section 5.); the previous parts deal with the corpus-based foundation of the work (section 2.), the classification of extracted collocations (section 3.) and the functional properties of the dictionary (section 4.).

The need for this study arose primarily from an analysis of the heterogeneous and largely inadequate treatment of collocations found in both general and special language dictionaries currently available on the market. A close investigation of these dictionaries, as

[1] This investigation is part of a PhD project at the Institute for Translation and Interpreting, University of Heidelberg (Germany).

DOI 10.1515/lexi.2011.016

well as of widespread online lexical databases, reveals four major orientations in the representation of collocations[2]:

1) Semasiological general language dictionaries that focus on the semantic properties of a given collocational element and, in most cases, include collocations in the phraselogical position of an entry (cf. Sabatini-Coletti 2007). The main drawback of this lexicographic model is that collocations, not being central to the lexical description of a lemma, are often not clearly distinguished. They can be found not only among phraseological expressions, but also among examples of free lexical combinations, and this aspect negatively affects dictionary usage efficiency.
2) Semasiological special language dictionaries that focus on the syntactic properties of collocations, which then function as the primary classificatory elements of an entry. These dictionaries of collocations use syntactic patterns as metadata for the further grouping of collocations, and they do so either implicitly (cf. BBI) or explicitly (cf. OCD).
3) Onomasiological dictionaries and lexical databases, in which single lexical items and collocations are placed into complex lexical networks (cf. WordNet). In WordNet, for instance, collocations are treated as independent entries, and therefore they belong to synsets (sets of synonymous lemmas)[3].
4) Lexical models based on the connection between syntax and semantics (cf. *Wortprofil im Französischen*, Blumenthal 2006); they are not concrete lexicographic works, but rather models of lexical analysis focusing on the interplay between the autonomous, intrinsic characteristics of a word, its combinatory profile and the way in which it conceptualises the world. This kind of analysis doesn't omit the diachronic perspective, since historical information about a word can be useful in comprehending its present lexical behaviour.

The aim of this corpus-based study has been to design a specific reference work for textual production and language skills improvement, one that contains clear lemmatisation principles, a transparent network of semantic relations both among single lexical items and among collocations, and a solid mediostructural architecture that enables the easy retrieval of exhaustive but also selective information during any query. Tarp (2008, 265) points out the great advantages of using the electronic medium together with conceptual structures when representing word combinations for text production[4].

---

[2] In this overview I have consulted Italian, English, French and Spanish dictionaries and lexical databases.

[3] In spite of its potential lexicographic effectiveness, this kind of resource still records only a small number of collocations. In WordNet collocations are defined as "string[s] of two or more words connected by spaces or hyphens" (http://wordnet.princeton.edu/man/wngloss.7WN.html, last visited: 13.05.2011). They are strongly idiomatic combinations with a high degree of morphological stability: examples would be *blue-collar* or phrasal verbs.

[4] Tarp (2008, 123) also introduces the idea of a new reference tool, the *leximat*, "consisting of a search engine with access to a database and/or the internet, enabling users with a specific type of communicative or cognitive need to gain access via active or passive searching to lexicographical data, from which they can extract the type of information required to cover their specific needs".

The case study presented in this paper focuses on a small set of nouns belonging to the semantic field PAURA (*fear/fright/worry*)[5]. The decision to work on a restricted number of words was based on the need to gain a thorough understanding of the syntactic and semantic behaviour of the lexical items and their collocations: such an in-depth analysis could only be conducted on the corpus occurrences of a small, homogeneous lexical sample. The semantic field PAURA has proven to be fertile research ground: there are, in fact, many lexical studies, including contrastive ones, dealing with this topic (cf., among others, Bergenholtz 1980; Wierzbicka 1998; Heringer 1999; Tissari 2007; Blumenthal 2002 and 2008; Rovere 2008). Moreover, the ontological status of emotions has been widely discussed within lexical semantics and is still a central issue in the construction of lexical databases that are based on semantic hierarchies. For instance, in the context of the project called WordNet Domains, in which domain tags are assigned to sets of lemmas, a great deal of attention has been devoted to the lexicon of emotions. WordNet-Affect in particular is a specific extension of WordNet Domains that is dedicated to labelling affective concepts (Valitutti, et al. 2004; Strapparava, et al. 2006).

Ontologies deal with questions concerning the existence of extralinguistic items such as physical objects, properties or actions, and with their categorization. An ontology is made up of hierarchical structures, such as taxonomies, as well as of formal descriptions of items and classes. According to the traditional, i.e., Aristotelian, view, items must fulfil certain necessary and sufficient conditions in order to belong to the same set with rigidly drawn category boundaries. Nevertheless, many concepts, or mental representations of natural items, are not easily definable, and it is only likely that their features will be present. Probabilistic models, which were developed in the fields of cognitive psychology and linguistics in the early 1970s, put forward an alternative approach to categorization, introducing the issue of probability with regard to categorial features.

Prototype theory is one instance of these probabilistic categorization models. At the core of this theory is the assumption that categorial membership is determined by a sufficient resemblance to prototypical items, i.e., the best representatives of a given class, and that categories are represented in terms of fuzzy sets of elements. In the cognitive sciences, prototypes have already been applied to basic human emotions, i.e., a primitive set of emotions, including fear, anger, happiness, sadness, and disgust, that have universal phenomenological or bodily components such as specific facial expressions (cf. Johnson-Laird/Oatley 1989; Ekman 1992).

A prototypical structure can easily be assimilated into a lexical hierarchy, thus becoming an integral part of an ontology. Prototypicality allows for more precise and explicit cross-referencing between a semantically central lexical nucleus (*paura*: the prototype) and the elements within its category. At the same time, *paura* is linked semantically to contiguous entities and concepts such as other basic emotions or secondary, i.e., complex, emotions. Many hierarchical structures that reflect semantic relationships among lexical items are of a taxonomic nature[6]. In taxonomies, subordinate elements inherit the features of the super-

---

[5] The English equivalents presented in this paper are obviously not meant to provide the reader with an exhaustive list of translation possibilities.

[6] Wierzbicka (1984) points out the co-existence of taxonomic and non-taxonomic supercategories. Whereas taxonomies label the relation "something is a kind of something else", non-taxonomic classification can apply to, for instance, partonomies, functional concepts (toys, weapons, etc),

ordinate node but also have some additional, specialized properties. The substantives belonging to the semantic field PAURA are linked to the prototype via specialization, i.e., a hypernym-hyponym relation (Y is a kind of X): *panico* (*panic*), *fobia* (*phobia*), *ansia* (*anxiety*), *orrore* (*horror*/*scare*), etc., are in fact subtypes of *paura*. *Panico*, for instance, shares the same general semantic features with *paura* ("intenso turbamento misto a preoccupazione ed inquietudine per qlco. di reale o di immaginario che è o sembra atto a produrre gravi danni o a costituire un pericolo attuale o futuro"[7], according to the first meaning of *paura* in Zingarelli), but, at the same time, it indicates a sudden, overpowering fear that causes irrational behaviour in individuals, groups of persons, or animals.

2. Corpus-based hypothesis testing

My initial hypothesis regarding the possibility of representing collocations within a syntactic and semantic framework has to be thoroughly tested and transformed into a concrete lexicographic description. It can be assumed that multiword expressions, just like simple lexical items, can be inserted into a lexical network in which they are linked to their constituent elements and to contiguous expressions. Contiguity, in this case, has to be understood in terms of the number of semantic, syntactic and pragmatic properties shared by linguistic objects. As collocations are the syntactic combination of two or more lexical items mostly belonging to different grammatical classes (NOUN + ADJ, NOUN + VERB, PREP + NOUN, etc.), they can be grouped together quite easily according to their internal structure (as we have seen, this is the typical form of collocation classification in the traditional dictionaries of collocations).

On the semantic level, contiguity extends beyond synonymy, and can be referred to other semantic relations like (co)hyponymy or partonomy. (Co)hyponymic relations play a major role in my taxonomic and prototype-based categorization, but synonymy is a more problematic issue. If we consider the specific contextual meanings of the substantival hyponyms of *paura*[8], we can generally observe that they stand for very similar concepts, yet they could be labelled, at the most, as quasi-synonyms. Pragmatic affinity, the last point, concerns the presence of similar usage patterns that depend on the contexts and situations in which sentences are placed. Pragmatically homogeneous collocations, for instance, share, on the diaphasic and diastratic levels, the same terminological field (*attacco di panico*/*panic attack* and *disturbo d'ansia*/*anxiety disorder* belong to the language of psychology and psychiatry) or the same register type (both *paura matta* and *paura del diavolo* indicate a great fear and are characteristic of colloquial language).

---

or collective concepts (an apple is not a kind of fruit and a table is not a kind of furniture: they are, rather, a variety of homogeneous kinds).

7 "an intense feeling of distress and worry caused by a real or imagined threat, which could possibly produce great damage or pose a present or future danger to us" [my translation].

8 Some of them are obviously polysemous words. The identification of different meanings should result in the placement of these substantives at different nodes of the lexical hierarchy since each parent or child node is linked to a specific semantic pattern.

My working hypothesis, of course, takes into account theoretical statements such as grammatical rules or pragmatic markers and assimilates them into an onomasiological pattern of collocation description. A corpus-based approach allows for the retrieval of a large amount of data on which this hypothesis can be empirically tested. Empirical verification consists of numerous steps involving both quantitative and qualitative analysis.

In the case of empirical data that does not fit the hypothesis, two solutions are possible: on the one hand, we can adjust the hypothesis, redefining its principles on the basis of corpus evidence; on the other hand, we can build corpus data into the aimed-at collocation description as empirically probable or perhaps marginal phenomena (Tognini-Bonelli 2001, 68–72[9]). In both cases, we have to assume that the degree to which inferences can legitimaltely be made from corpus data strictly depends on the intrinsic characteristics of the corpus itself, namely on its size and linguistic properties. The way in which corpora are edited for lexicographic purposes should be based on fundamental requirements such as the type of dictionary, its genuine function, or its potential users. Despite these considerations and the level of accuracy in preparing the text and metatext, corpora always constitute open inventories of occurrences, the representativeness of which is not absolute, but can only be judged in relation to their final purpose.

In this case study, collocation patterns have been derived from two sources. I have used a lexicographic corpus made up of the most common general language Italian dictionaries and some dictionaries and encyclopaedias of psychology and philosophy. These dictionaries contain differentiated data on the level of linguistic variability and are generally characterised by synoptic structures and a highly synthetic linguistic and metalinguistic description. For these reasons, they not only constitute an essential data source, they also provide useful microstructural models. Nevertheless, because of its synthetic and systematic aim, the lexicographic corpus inevitably turns out to be incomplete from the perspective of a realistic reproduction of natural communicative situations. The need for a description of language in authentic use leads us to the choice of a non-lexicographic data source, namely, an electronic corpus that contains texts from major contemporary Italian newspapers and covers a period of seven years. Newspaper articles constitute a rich collection of authentic language patterns in context, and their syntactic, semantic and pragmatic features can be evaluated directly in textual segments of varying structure and length. Despite its apparent homogeneity, the newspaper corpus is extremely heterogeneous on the level of text typology, since each article belongs to a specific newspaper section (news, editorials, reviews, etc.) with its own writing conventions. In any case, narrative and argumentative articles can be regarded as the dominant text form. A high degree of diamesic variation, i.e., the alternation of written and spoken language, for example in direct and reported speech, can also be observed. At the same time, peculiar newspaper styles and idiolectal patterns used by journalists play a role in determining such typical linguistic fragmentation.

Words indicating emotions, and therefore the nouns belonging to the semantic field PAURA, are quite frequent in newspaper texts. The delivery of information is generally coupled with a high degree of expressiveness, which can be observed both in the lexical choice and in the emphasis on connotative meaning through the recurrent use of stylistic

---

[9] A third possibility proposed by Tognini-Bonelli is so-called *insulation*, which consists in keeping corpus data and theory apart. I'm not incorporating this method into my study, since it would not suit my project's lexicographic perspective.

techniques such as ellipses, antitheses, metaphors, or intertextual reference. Moreover, a large spectrum of specialized terms concerning the fields of, for instance, politics, economics, law, or science, is available in the newspaper corpus. On the syntactic level, emotional writing is achieved through the frequent use of nominal phrases and short sentences, which often resemble the typical style of headlines, as well as through highly marked structures: subject-verb inversions (*cresce la paura di qc/fear of sth is growing*), *frasi scisse/pseudoscisse/specificative*[10] (for the latter type: *la paura è che.../one's fear is that...*), passive and absolute participle constructions (*la paura è passata/the fear is over*, *passata la paura.../(once) fear is over...*), the use of the form *c'è* at the beginning of a sentence (*c'è paura/there are fears*), and hyperbolic or exclamatory sentences (*niente paura*/*don't worry*, *no worries*).

Collocative combinations available in the lexicographic corpus are integrated into the data set extracted from the newspaper corpus. From a methodological standpoint, this study is based on the assumption that the final lexicographic data depends completely on the properties of the corpora. Precisely for this reason, no other collocations have been added for the moment to those found in the corpora: the inclusion of numerous lexical co-occurrences would be justified at least by intuition[11], but this wouldn't be a reliable method, since it could easily lead to arbitrary choices. Throughout the study, this procedure will allow for better control over the large amounts of corpus data available. Obviously, from another methodological perspective, the original sample data could also be reduced to a more manageable size and at the same time be supplemented with missing elements in order to fit a different and probably more specific lexicographic function. Even though I have clearly defined a function and a potential user for the dictionary, the main goal of this study is not so much to design a reference work in its final form as it is to model an overall lexicographic structure suitable for the treatment of collocations.

The statistical significance of the extracted co-occurrences is measured by means of the log-likelihood function, a useful means of detecting sparse data and rejecting semantically general lexical items. Automatically retrieved information such as likelihood ratios, frequency values, positional features and part-of-speech tags identify what I call *collocation candidates* and provide initial indications of how they should be evaluated. In the next phase of the study, non-automatic evaluation allow lexicographically irrelevant phenomena to be excluded; these include spelling mistakes, newspaper section names, idiolectal and intertextual utterances, and words exclusively related to occasional events. Even corpus-specific properties like high likelihood or frequency values for words accidentally co-occurring with *paura* and other nouns, as well as significant likelihood values for words that are not proper collocation partners of a given noun but rather belong to one of its stereotypical scenes can be identified at this stage. Non-automatic evaluation also aims at the detection

---

[10] These kinds of structures aim in different ways at the focalisation of the nominal element of the sentence, here the words indicating emotions.

[11] For instance, this is the case for verbal patterns, such as modes and tenses, that are not present and for missing plural or singular forms of nouns and adjectives. But this can also happen in the case of word combinations in the form of collocations that are certainly part of our mother-tongue vocabulary although they are not recorded in the corpora.

of important linguistic aspects the statistical measures cannot account for, like complex predicates, non-binary combinations, and highly phraseological expressions[12].

While these steps are being implemented, the definition of collocation on which this work is based must be kept at the forefront. I define a collocation functionally either as an idiomatic, multiword expression subject to restricted compositionality, substitutability, and modifiability, or as a familiar word combination recurring in our mental lexicon, mostly in association with typical scenes, i.e., complex scenarios depicting a certain situation, and connected to well-defined linguistic frames (Conway/Bekerian 1987; Fillmore 1977). This twofold, lexicographically oriented specification is an indispensable condition for an adequate treatment of the extracted collocation candidates. The proposed dictionary is intended to be a resource for text production, and as such, it should provide dictionary users with a wide range of phraseological occurrences together with an appropriate macro- and microstructural descriptive framework. From a lexicographic perspective, drawing strict boundaries between different levels of phraseological cohesion such as collocations, semi-idiomatic and idiomatic expressions[13] is likely to be counterproductive, since it could possibly lead to arbitrary classifications resulting in obvious mistakes in the act of dictionary usage[14]. Despite the presence of razor-sharp definitions that, on the theoretical level, strive to distinguish collocations on the one hand from metaphors and idiomatic expressions on the other, in concrete language situations we can often observe the overlapping of definitionally based features belonging both to collocations and to other multiword co-occurrences. These "gradual transitions" from one phraseological form to another (Tarp 2008, 254–255) are therefore a common linguistic phenomenon; for this reason, the class of semi-idiomatic and idiomatic expressions should be included in the dictionary without strict classification, whereas collocations as a whole should be separated from the class of free lexical combinations[15].

In this way, the initial hypothesis proves to be realistic: the extracted co-occurrences, the so-called collocation candidates, are judged on the basis of their lexicographic relevance and integrated into a lexical network. Whenever a selected collocation doesn't fit into the pre-built classification, the above-mentioned solutions, i.e., the restatement of the classification pattern and the ranking of some data as corpus-dependent phenomena, are applied in sequence. Thanks to this stepwise procedure, the end result is a coherent and homogeneous method of collocation description.

---

[12] Morphosyntactic fixedness can be determined by means of conventional tests involving substitutions (pronominalisation, anaphora, interrogative or relative clause) and modifications (adjectival modifier, plural form).

[13] According, among others, to the classification of phraseological units proposed by Cowie 1983 and 1998.

[14] According to Tarp (2008, 250–254), a typological subdivision of phraseological units doesn't take into account the actual needs of users, especially learners. The main goal of structuring word combinations in a dictionary is not so much to reflect linguistic principles but rather to enable the dictionary user to find specific data in the most efficient way. Tarp obviously distinguishes the needs of learners at the beginner level from those of intermediate and advanced learners. In our case, the potential users are professional translators with a high level of language proficiency, who need to use metaphoric and non-metaphoric idioms in order solve specific translation and text production problems in a subtle and flexible way.

[15] On the question of whether theoretical issues are generally relevant to practical lexicographic use, cf. Tarp (2008, 249–250).

3. Collocation classification and article-internal positioning

Pre-dictionary analysis involves, on the one hand, a detailed description of the different syntagmatic frames in which *paura* and the other nouns may occur. During a first comparison of the textual realisations of the selected nouns, new details emerged about the prototypical role of *paura* within the semantic field. We observed that the prototype imposes on its lexical group a specific syntactic behaviour that takes the form of a general framework of structural patterns. A new level of specialisation follows in the transition to the lower nodes of the hierarchy, where the syntactic properties of the prototype/hypernym are adopted and adjusted to the linguistic characteristics of the hyponyms. If we take into consideration both the Fregean view regarding the saturated or unsaturated nature of words and Lyons' theory about the ontological classification of lexical entities, *paura* and the other nouns can be said to be unsaturated second-order entities with no complete, independent meaning. As second-order entities, they indicate events, processes, and states of affairs (Lyons 1977, 443), mostly depending on the aspectual and actional features of a specific context. They often build complex predicates by joining a rather homogeneous set of support verbs (for instance, *avere/provare/sentire/avvertire*, indicating the act of experiencing the emotion, or *fare/mettere/incutere/provocare*, indicating the act of triggering it), which are also shared by words indicating other types of emotions. Moreover, their meaning must necessarily be specified and completed, within an argument structure, by collocation partners, i.e., all the lexical items to which they are combined within a collocation.

On the microstructural level of the dictionary, the syntactic and semantic dimensions are deeply interwoven, since differences in word and collocation meanings often strictly depend on differences in syntagmatic structures. For this reason, an initial syntagmatic description is essential. On the other hand, pre-dictionary work also includes an investigation of the semantic features of collocation partners as well as their classification into homogeneous groups. The corresponding procedure involves a series of steps, through which available collocational data are progressively classified according to the final lexicographic aim. During classification, more and more subtle analytic parameters are introduced as notational tools.

The first step in classification depends strongly on the syntactic (part-of-speech) function of the collocation partners occurring with the selected nouns. Co-occurrences are grouped into easily detectable units, i.e., *paura* + ADJ, *paura* + NOUN, *paura* + VERB, *paura* + PREP, *paura* + exclamatory ADJ and complex phraseological expressions containing *paura* and varying lexical elements. Within each of these groups, the collocation partners of the reference noun are sorted into functional semantic patterns with the help of the above mentioned analytic parameters, choosing the best description model for each grammatical class. Adjectival collocation partners are primarily identified through a combination of thematic relations and a framework of principles and constraints derived from psychological studies, mostly concerning the origin, quality, intensity, duration, and adequacy of emotions. Thematic relations are most adequate for representing the semantic roles of the subjects involved in the emotional events, independent of their syntactic placement. The key thematic roles in the case of *paura* and its semantic field are the Experiencer, i.e., the entity that receives an emotional input, and the Agent/Cause, i.e., the entity that deliberately or accidentally triggers an emotion. Together with other thematic relations, they also

play a part in the identification of substantival collocation partners[16]. Experiencer and Agent/Cause can be seen as mirror-image roles, which best describe the logical core of the emotional event of fear. Typical examples for their adjectival and substantival realisations in connection with *paura* are: *paura collettiva/della società* (*collective/social fear*) and *paure infantili/dei bambini* (*children's fears*) with an Experiencer role taken by the respective collocation partner, or *paura inflazionistica/dell'inflazione* (*inflation fear*, *fear of inflation*) and *paura comunista/del comunismo* (*communist fear*, *fear of communism*) with a Cause role. Despite the frequent correspondence between different grammatical classes, in the majority of cases a thematic relation connected to a specific semantic field shows a clear preference for one part of speech over another. For instance, the social, political, or economic origin of *paura* and the other nouns is best realised by adjectives (*paura economica/nucleare/razziale*[17]), whereas substantives are the most common form for emotion-triggering natural phenomena or abstract entities (*paura del buio/del vuoto/degli animali/del male/del futuro*[18]).

It is not always possible to assign a thematic role to a collocation partner, such as in the case of qualitative adjectives like *grande/profonda/improvvisa/segreta/legittima/vera paura* (*great/profound/sudden/secret/legitimate/true fear*), which give information about the quality of the noun they modify, prepositional phrases like *attimi di paura* (*moments of fear*), in which *paura* serves as the prepositional complement governed by another noun, or coordinate structures following the pattern NOUN + NOUN(S) like *paura ed angoscia* (*fear and anxiety*) or *dolore*, *rabbia e paura* (*pain, rage and fear*). The thematic properties of these phenomena cannot be evaluated without considering a concrete contextual environment.

The classification of verbal collocation partners is based on the interplay of the grammatical function of the verbal arguments as subjects, direct objects or prepositional complements, the thematic role assigned to subjects other than the analysed noun, and verbal aktionsart, which also affects verbal aspect. The remaining structures, namely *paura* + PREP (*senza paura*, *without fear*), *paura* + exclamatory ADJ (*che paura!*, *what a fright!*), as well as complex phraseological expressions (*la paura è una cattiva consigliera*, *fear is a bad counselor*) are listed separately and require no internal subdivision for the moment.

The initial classification of the selected data provides a comprehensive overview of the available collocations and offers a starting point for developing a more refined strategy of lexicographic description. This involves the reorganisation of the collocational data into five major classes with a targeted clustering of the above-mentioned classification parameters[19]. The five classes fall under the microstructural layout of the dictionary and occupy the particular position of the lexicographic entry that is dedicated to the treatment of col-

---

16 We assume that, from a lexicographic, not purely theoretical perspective, and taking into consideration the large amounts of collocational data which have to be classified, the distinction of too many thematic relations should be avoided. In fact, it is not always possible to trace clear boundaries between similar roles, and overspecification might produce unfavourable effects on lexical and lexicographic description.

17 *Economic/nuclear/racial fear*.

18 *Fear of the dark/of the void/of animals/of evil/of the future*.

19 For example, adjectival collocation partners no longer belong as a whole to a single group; rather, qualitative adjectives have been detached from adjectives linked to a thematic role, which are now entered in Class 1 together with substantives.

locations (see below). Each class is identified by a general syntactic pattern and subsequent syntactic and semantic specifications, whose aim is the isolation of homogeneous, easily distinguishable sets of collocations[20]. For instance, *paura dei ragni* (*fear of spiders*), which belongs to the first class, can be found by following this access route:

CLASS 1: *paura* + PP (prepositional complement) → Agent/Cause of *paura* → people, personified entities, animals → *paura degli animali* (*fear of animals*), *paura dei ragni* (*fear of spiders*), *paura dei serpenti* (*fear of snakes*), etc.

I will now illustrate the access structures of the other collocational classes of *paura*:

CLASS 2: *paura* + AP (quality) → pathological origin → *paura patologica* (*pathological fear*), *paura fobica* (*phobic fear*), *paura notturna* (*nocturnal fear*), *paura ossessiva* (*obsessive fear*), etc.

CLASS 3: *paura* + V → *paura* (subject) + V → continuative verb → *la paura dilaga/serpeggia/si diffonde* (*fear spreads*), etc.

CLASS 4: *paura* within a PP → N + *paura* within a PP (prepositional complement) → *momento di paura* (*moment of fear*), *ore di paura* (*hours of fear*), *notte di paura* (*night of fear*), etc.

CLASS 5: *paura* within other structures → *da paura* (*terrible*, *incredible*, *fantastic*), *che paura!* (*what a fright!*), *occhi pieni di paura* (*eyes full of fear*), etc.

The last class is a collection of all those combinations that cannot be fitted into the other four syntagmatic frames; in most cases they are non-binary idiomatic expressions. Obviously, phraseological combinations with a high degree of lexical fixedness and restricted compositionality can be also found in the other classes and are explicitly marked as such by pragmatic labels indicating their complete or partial figurativeness. The syntax-semantics interface is in fact completed by a pragmatic layer, in which collocations can be assigned one or more markers specifying register, style, or terminological information when required.

This final classification is solely lexicographically oriented and should therefore be preceded by an intermediate stage, in which every collocation is coupled to a formal description, so that it can be inserted into a lexical database serving as a data source for the dictionary. A comprehensive formal description of collocations should include, on one side, clear cross-referencing a) between a word and semantically related words, b) between a word and its collocations as well as c) between collocation partners belonging to the same multiword expression, and, on the other side, structural information in the form of syntagmatic, syntactic and semantic tags. These tags label, respectively, the phrase structure of the collocation, the grammatical function of verbal arguments, the thematic role of the collocation elements, and, if required, other semantic features.

As we have seen, all of the collocation partners of a certain noun are represented within a functional hierarchy of metalinguistic items which makes it possible for the dictionary user to look up an entry and clearly identify uniform groups of collocations. The microstructure of an entry is based on a lemma sign, a comment on its form, a paraphrase of its

---

[20] The potential user is assumed to possess sufficient usage skills to be able to cope with relatively complex metalexicographic structures. Yet this lexicographic model could be modified and adapted to learners' needs by simplifying metadata.

meaning and the interplay of two distinct modules, each made up of various positions (figure 1[21]). One module is specifically dedicated to the treatment of collocations and contains the five classes that were used to describe the collocational patterns above. The other module relates to the position of a given lemma in the lexical network that constitutes the macrostructural layout of the dictionary: in the entry, each lemma is displayed as a node belonging to the underlying taxonomic tree. Of course, for space and clarity reasons, the dictionary user is presented with a relatively small section of the taxonomy at once, yet this section has to be large enough to include at least the immediate hypernyms and hyponyms of the lemma node. A further specification of the lemma's semantic relations, i.e., a sort of "zooming out" of the lexical network, is possible at any time by following reference-related links.

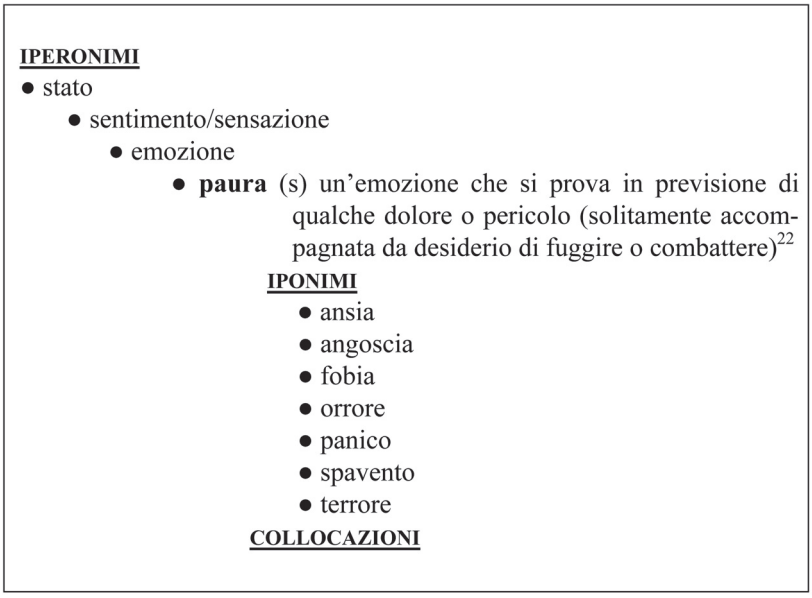

Figure 1 – Microstructure of the entry *paura*

The biggest part of the items composing our mental lexicon are collocations and, in general, the familiar association of words. In order to reproduce, as realistically as possible, the way in which our minds organise extralinguistic knowledge and the corresponding lexicon, I have chosen to take advantage of the electronic medium and of the onomasiological arrangement. If we leave aside the obvious case of WordNet and the many lexical tools derived from it, which are based strictly on an onomasiological approach, there are still some other online lexicographic resources providing information on semantic fields, like the Words-

---

21 Taxonomic labels are a revised version of the ones found in the WordNet and in the Italian MultiWordNet hierarchies.

22 "An emotion experienced in anticipation of some specific pain or danger (usually accompanied by a desire to flee or flight)" (WordNet).

myth thesaurus, the OneLook metadictionary, or the Lexical FreeNet thesaurus. But in practice, the lack of a systematic onomasiological approach produces only a partial, rather useless data representation. Moreover, collocations are not integrated into the onomasiological description.

An onomasiological dictionary leads the user from an ontological category or a semantic field to actual lexical elements, which are grouped into homogeneous conceptual areas. Onomasiology provides the method of arrangement governing the lexicographic macrostructure, but it also prevails on the semasiological order on the microstructural level, in which a semantic, conceptual classification appears to be better suited for the treatment of collocations than plain alphabetical listings[23]. The onomasiological form of arrangement itself has some general disadvantages, like the subjectivity of its categorization (Bergenholtz 1980, 137) and its unsuitability in representing lexicographic data for text reception (Baldinger 1998, 2124). But it fosters the space-efficient treatment of data, allowing for a functional clustering of syntactically and semantically contiguous co-occurrences.

4. The functional properties of a dictionary of collocations

The genuine purpose of the proposed electronic dictionary is the onomasiological representation of Italian collocations belonging to the semantic field PAURA. From a typological perspective, this reference work can be described as a monoinformative[24] dictionary that is focused on syntagmatic structures and which systematically integrates the partial collocational representation in monolingual and bilingual general language dictionaries with more comprehensive data. The surface syntagmatic description is combined with an underlying paradigmatic approach to lemmas and collocations. Semantic relations, like hypernymy or synonymy, among lemmas as well as among collocations can be seen in terms of paradigmatic structures: within a specific context-based collocation, for instance, a collocation partner can often be replaced by a semantically contiguous lexical element such as a co-hyponym (*terrore/orrore della solitudine*, *terror/horror of loneliness*) or quasi-synonym (*ansia/angoscia da separazione*, *separation anxiety*), but also by a generalising hypernym or a specialising hyponym (*gettare nella paura/nel panico*, *send into panic*)[25]. This aspect concerns the extent to which a collocation is lexically flexible, i.e., the degree of substitutability of its elements. In text production situations as well as in the preliminary, reception stage of translation, the act of paraphrasing a text for comprehension purposes

---

[23] Nevertheless, alphabetical order could play a role in the final arrangement of the conceptually organised data. This would undoubtedly increase the functionality and user-friendliness of the dictionary.

[24] The core of lexicographic treatment is in fact a single lexical phenomenon, namely collocation. Yet, some other linguistic elements play a central part in lexicographic data processing, for instance, semantic relations on the taxonomic level.

[25] In the majority of cases, the exchange of collocation partners seems to take place on the hypernymic-hyponymic level. This fact strongly supports the view that the prototypical lexeme has a significant syntactic and semantic influence on the way in which the members of its lexical group collocate with other lexical items.

or for the targeted production of coreferential expressions typically implies the application of these strategies. These interference effects between collocations and, as a consequence, the varying degree of paradigmatic interchangeability of their constituent elements should obviously be displayed to the dictionary user as a component of collocation description. A clear and coherent cross-reference system is necessary to fulfil this purpose.

This dictionary acts as a resource for text production in Italian as a mother tongue or as a second or a foreign language, and as such it is primarily intended for communicative, not cognitive, purposes. Since it is a monolingual dictionary, there is no need for it to include interlinguistic or intercultural information. Even the item giving the meaning of a lemma should be restricted to a few remarks, and its main purpose should be to provide the reader with general semantic information about a word and about its taxonomic environment. Providing a detailed commentary on the meaning of a word should be one of the tasks of a general language dictionary, not of a dictionary of collocations. Reference works with different purposes should complement each other in each usage situation. The possible drawbacks of cumulative data arrangement should nevertheless be considered and avoided without resorting to standard lexicographic definitions. Without passing judgment on the general utility of cumulative dictionaries, which are a valuable reference source for text production and for advanced users (Wiegand 1999a and 1999b), I would like to point out the problem of presenting the dictionary user with a large amount of undistinguished data extracted from the corpus. At the same time, at least overall semantic indications should be provided for not entirely transparent combinations. For that reason, I consider the ontological categories and subcategories that I used in the final classification of collocations to be a kind of general semantic marker. Moreover, the prototype approach indirectly influences the semantic labelling of collocations by highlighting the connection between the prototypical lemma (with its hypernymic meaning) and the collocations of its hyponyms.

A dictionary, in accordance with its genuine purpose, can be consulted either during an isolated, specific act of usage or while performing a reading act with no predefined objective (Wiegand 2008, 9–10). An active dictionary of collocations should enable the user to a) activate passive knowledge (for instance, to find a collocation starting from a collocation partner), b) prove the correctness of a hypothesis (for instance, about the existence of a certain collocation or about a collocation having a particular pragmatic feature), and c) retrieve new collocational or metacollocational knowledge.

We will now look more closely at the specific user needs that the dictionary is intended to meet. In comparison with paper dictionaries, the electronic medium makes it possible to perform more accurate search queries and is therefore suitable in numerous usage situations. For instance, the dictionary user might want to find

- all collocations of a given noun,
- all nouns combining with a given collocation partner, or
- collocations of a given noun according to one or more predefined parameters, such as the grammatical or lexical character of a collocation, part of speech, thematic relation, onomasiological (ontological) placement of a collocation partner in the taxonomic hierarchy, the grammatical function of the noun, the pragmatic features of the collocation, or its idiomatic properties. In these examples, the search query starts from a single lexical element, i.e., one of the selected nouns. The first two tasks could potentially also be performed by paper dictionaries, as long as they record collocations both under the

basis and under the collocator[26]. The electronic medium opens up additional possibilities, enabling the dictionary user to discover, through a detailed cross-referencing system, the prototypical role of a certain lexical element as well as similarities between collocations inside taxonomically contiguous semantic fields[27].

5. Mediostructural properties and dictionary consultation

5.1 The cross-referencing system

From a mediostructural perspective, every type of dictionary should be provided with a cross-referencing system capable of meeting specific user needs: cross-reference conditions depend strictly on the object, the form, and the function of a dictionary (Reichmann 1991, 1063). In accordance with the primary function of the dictionary described in this study, the majority of the cross-references have a text-production and a translation-supporting aim. On the level of the lexicographic object, they can be, for example, collocational, hyponymic, hypernymic, or synonymic. The term *cross-reference* as used here does not refer to actual dictionary-internal and consequently article-internal text segments. With the aid of a specific item or text segment enabling cross-reference, dictionary users come to know that they are required (or at least advised) to perform a cross-reference follow-up act in order to reach a specific cross-reference address. In this sense, a cross-reference doesn't belong to the dictionary-internal data but rather is part of the lexicographic information that can be cognitively retrieved by using well-defined functional indications (Wiegand 2002, 179–180, 211). In the proposed dictionary of collocations, the cross-reference address, i.e., the access address to which the cross-reference leads[28], is characterised by an internal access structure (or internal address).

The electronic medium can significantly enhance the mediostructural quality of the dictionary, ensuring data consistency and transparency of cross-referencing. Kammerer (1998, 4–6) considers hyperlinks to be a relation, i.e., a set of sorted pairs that satisfy the following condition: given x and y as elements of the set of informational units (*Menge der informationellen Einheiten*)[29], x is connected to y by a specific relation. A hyperlink is a function and can be described as a tuple combining a cross-linking indicator

---

[26] In this study, I generally reject the use of the terms *basis* and *collocator* as defined, for instance, in Hausmann (1989, 1010; 2003, 315), Bartsch (2004, 32–38), and Schafroth (2003, 400), since they do not always have a clear practical use in describing the constituent elements of a collocation. The identification of the basis, referred to as the semantically autonomous, co-creative part of the collocation, appears to be especially problematic in the presence of non-binary collocations and of co-occurring substantives like *terrore della morte* (*terror of death*), *brivido di paura* (*quiver of fear*), and *film dell'orrore* (*horror film*).

[27] For example, fields indicating primary emotions such as fear, anger, joy, sadness, or disgust, which belong to the same taxonomic level and can thus be considered co-hyponymic nodes.

[28] This should be distinguished from the article-internal reference address from which the cross-reference follow-up procedure starts.

[29] On the problematic term *informationelle Einheit* introduced by Kuhlen, cf. Kammerer (1998, 4).

(*Verknüpfungsanzeiger*) and an item giving the cross-reference address (*Adressenangabe*). The idea of an existing lexical network has the advantage of supporting this formal combination of element pairs: on the one hand, the electronic, addressed source data and on the other hand, electronic, addressed target data (Müller-Spitzer 2007, 146–147).

Within an onomasiological dictionary of collocations, cross-references should support both the taxonomic system on which the lexical network is based and the interference effects between collocations of different lemmas. In this dictionary, hyperlink targets are always elements of the lexicographic object and thus correspond to the intrinsic, formal, and functional properties of the dictionary. Every cross-reference relation originates in fact in a hyperlink connecting a single word, i.e., a taxonomic node, or word combination to another. We can speak of a bidirectional cross-linking process, since it can be gone through at any time in the inverse direction. Subsequent cross-reference follow-ups are also possible; starting from the addressed informational unit, we can have a further cross-reference branching.

Figure 2 – Taxonomic cross-reference

Within each dictionary entry, cross-referencing on the taxonomic level enables the user to discover the immediate hypernyms and hyponyms of a given lemma. Each lexical item is an active element linked to other isomorphic data clusters. Starting, for instance, at figure 1 and following the hyperlink to the word/concept *emozione*, we obtain the results shown in figure 2, in which *emozione* is a lemma sign placed in the centre of the corresponding entry. Lexical items such as hypernyms could also be displayed as vertical drop-down menus and be connected in turn to other superordinate and subordinate elements. In this case, cross-reference items are implicit, since they are not accompanied by a specific item giving a cross-reference relation (*Verweisbeziehungsangabe*, cf. Kammerer 1998, 3); they are simply made up of lexical items that can be activated by the user in order to reach a certain target, without the cross-linking indicator being immediately visible. A progressive visualization of the taxonomic nodes inevitably produces information redundancy, since each cross-reference follow-up procedure leads to a partial recurrence of the previous results. What would appear to be a negative aspect of consulting a dictionary turns out to be an advantage of this lexicographic model, since it systematically presents the user with a coherent overview of the lexical network, without interfering with the rapidity of usage.

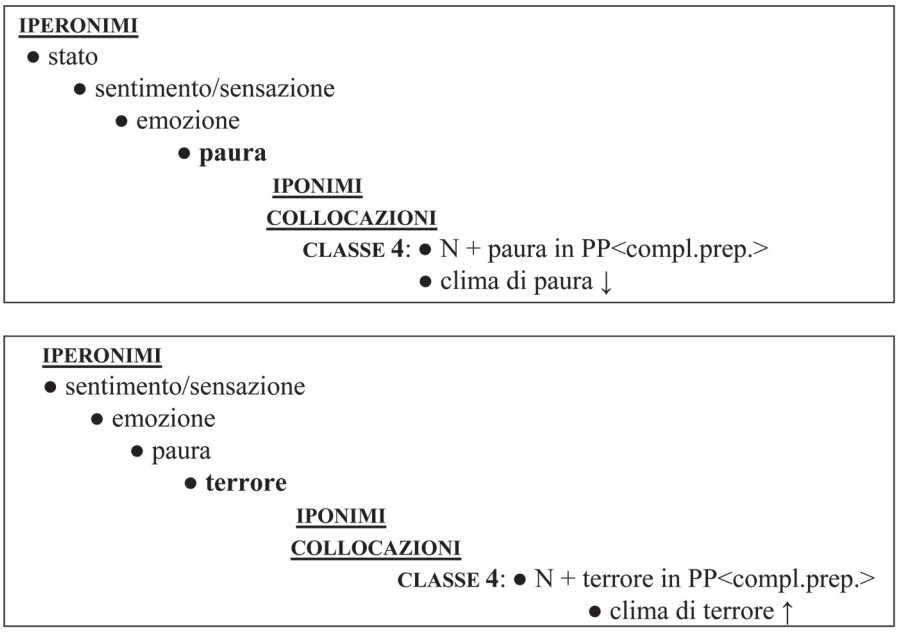

Figure 3 – Cross-reference between collocations: downward hereditariness

On the level of collocations, the goal of cross-reference is to point out which word combinations share at least one component and in which direction collocation partners are transferred from one combination to another inside the semantic hierarchy (figure 3). Cross-reference items are now explicit: each collocation that derives an element from one of its hypernyms or hyponyms is signalled by an upward or downward arrow. In this case the arrow is an item giving both a cross-reference relation and the origin of this relation[30], whereas the cross-linking indicator is given by the entire collocation (*clima di paura/atmosphere of fear*). The cross-reference target is a similar collocation (*clima di terrore/atmosphere of terror*) inside the entry of the corresponding hypernym or, here, hyponym (*terrore/terror*).

This kind of cross-reference information brings with it a clear representation of the semantic relations among lexical elements, but most of all, it allows for a transparent and more precise motivation of collocational meaning and provides the necessary lexical tools for producing collocational variation on the paradigmatic level. Although it is difficult to generalise the phenomenon of collocational hereditariness, a phenomenon that should be further investigated with more appropriate tools, I have observed that, in the majority of cases, the hereditariness of collocation partners follows a downward taxonomic path, i.e., hyponyms are more likely to inherit one or more collocation partners from the prototypical lexeme or, in general, from their hyponym. One exception, motivated by the influence of

[30] Since this second purpose is the most relevant one, arrows should only be placed after the cross-reference target (cf. figure 7). It would be quite useless and disruptive to have it also after the item giving the cross-reference address.

special language on general language, is exemplified by *angosce notturne* (*nocturnal anxiety attacks*), a word combination that belongs to psychological and psychiatric terminology. Even though *angoscia*, in its specialised meaning, doesn't belong to the hyponyms of *paura*, the collocation partner *notturno* can be transferred to *paura* (*nocturnal fears*). The underlying assumption is that this transfer is supported by the non-specialised meaning of *angoscia*, which functions as a semantic mediator (figure 4)[31].

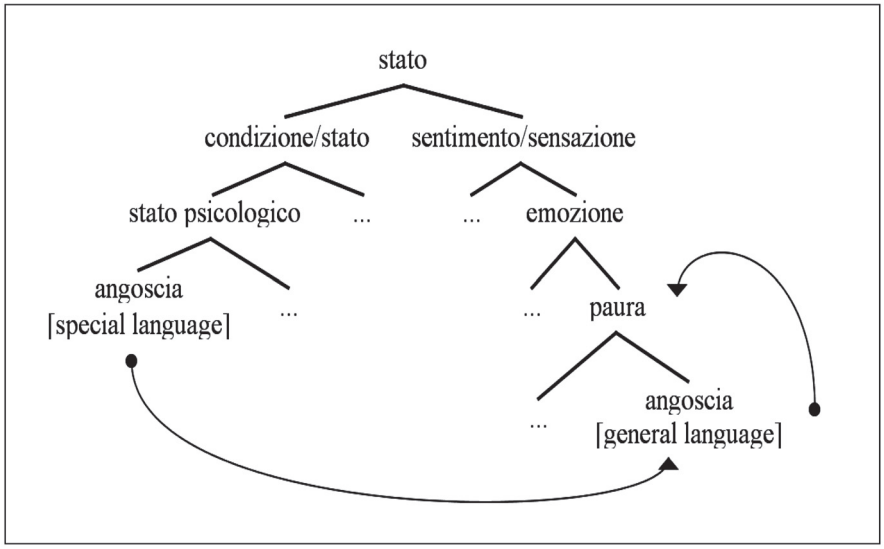

Figure 4 – Upward hereditariness of collocation partners

A less common kind of cross-reference relation involving collocations is the sharing of a collocation partner between co-hyponyms, without the collocation partner being inherited in the first instance from their common hyponym. For example, *ansia da separazione* (*separation anxiety*) and *angoscia da separazione* (*separation anxiety*) fall into this category. In this case, the item indicating the cross-reference relation might be a horizontal double-headed arrow (↔), which has the advantage of graphically recalling the location of the two lemma nodes (*ansia* and *angoscia*) on the same taxonomic level.

Another lexicographically interesting, though not very frequent, phenomenon we can observe about the selected collocations is the presence of combinations that are quasi-isomorphic as regards their collocation partners, and which are quite similar in their syntactic and semantic properties. Each time these combinations have to be assigned to different categorial classes (cf. section 3.) and thus entered into different article positions, it could

[31] In the semantic hierarchy that is at the core of the lexical network, all nodes referring to specialised concepts have to undergo a rigorous, scientifically based process of categorisation. Here the specialized fields involved are psychology, psychiatry, and philosophy, and in the case of the sub-branches of the node *condizione/stato* (*condition/status*), the theoretical source of categorisation is the latest edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM 4), published by the American Psychiatric Association.

be helpful to provide the dictionary user with items highlighting such close resemblance. From a methodological perspective, the diverging classification of these collocations can mainly be ascribed to their differing syntagmatic patterns; as we have seen before in this study, syntagmatic patterns are the first criterion for class identification. This phenomenon can be now illustrated by comparing the word constructions

a) *qu trema per la paura* (*sb trembles with fear*) and
b) *a qu tremano le gambe per la paura* (*sb's legs shake with fear*).

On the lexical (collocational) level, they have the substantive *paura* and the verb *tremare* (*to tremble*, *quiver*) in common, producing a quasi-isomorphism. Their syntactic structure mainly differs in the grammatical subject of *tremare*. From a semantic point of view, the second combination specifies the meaning of (a) by introducing a concrete physical element into its abstract stereotypical scene. There are no pragmatic differences between the two collocations. Despite their overall affinity, collocation (a) has been assigned to Class 3, whereas (b) has been entered in Class 5. This choice guarantees the syntagmatic homogeneity of collocational data within each of the five classes. In order to offset the possible lack of homogeneity in the representation of collocational variance, a further item giving the cross-reference relation between (a) and (b) has to be introduced. The item signalising the presence of quasi-equivalent collocations is ≈ (approximate equality).

5.2 Lexicographic search procedures

After outlining the main tasks performed by a dictionary of collocations in addressing special user needs (cf. section 4.), I will now expand on the issue of search procedures and search options. At this point, it is useful to draw a comparison between some online electronic dictionaries that record phraseological word combinations and to observe the search procedures they support. Quite often, as in the case of the online OCD or the Cambridge International Dictionary of Idioms, online lexicographic resources are the result of the hypertextualisation of print dictionaries and, as such, cannot exploit the advantages of the electronic medium in a profitable way. They usually allow a simple search query, i.e., the search for a single lexical item, whose output is a lemma and its dictionary entry, mostly without hyperlinks. The Free Dictionary is another online resource that retrieves its data from paper dictionaries[32] but which enhances data consultation by means of a full-text search option and the ability to search for word combinations beginning or ending with a specific lexeme. The Danish Ordbog over Faste Vendinger, with both a print and an electronic version, subdivides lexicographic information according to the type of communicative situation in which the dictionary is used; moreover, phraseological expressions are represented in a semantic network connecting one expression to another by means of mental associations.

In the French BLF, word combinations (collocations, idioms, proverbs) and their equivalents in five languages can be found by entering either a single word or two exact col-

---

[32] The lexicographic basis of this dictionary is the *Cambridge International Dictionary of Idioms* as well as the *Cambridge Dictionary of American Idioms*.

location partners into the search box. The search results are unfortunately characterised by a great degree of redundancy, since each word combination is accompanied by the definition of its elements. The DiCouèbe, a user interface that retrieves lexicographic data from the French DiCo database and presents them by means of interesting paradigmatic and syntagmatic connections, focuses mainly on the presentation of data, and as a consequence, queries and access paths require a high level of expert knowledge both of the lexical functions and of the underlying Meaning-Text Theory[33].

The Spanish DiCE is a dictionary of collocations that is also based on Mel'cuk and Polguère's theory of lexical functions, but with metalinguistic explanations supporting the search options and a slight increase in user-friendliness. In the advanced search options the three elements *lema*, *funcion léxica*, and *valor* can be combined in different ways to obtain specific results. The simple search option leads the user to a single dictionary article, in which collocations of a given lemma are grouped according to the lexical function of each collocation partner and are followed by contextualised examples. Hypertext links allow for improved data distribution and visualization.

In these examples, with the exception of the DiCE, the search output, regardless of the options available, mainly consists of unselected or poorly differentiated clusters of data; these compel the user to perform a new, time-consuming, and often ineffective manual search. This kind of data presentation offers a fragmentary view of linguistic items instead of highlighting their syntactic and semantic contiguity. To achieve better accessibility, all of the data and metadata that are part of the lexicographic object of an electronic dictionary should be retrievable by the user via a search query. The linguistic items belonging to the lexicographic microstructure should be regarded as simple or combinable search objects, whereas metalinguistic items should serve as search filters. In the proposed dictionary of collocations, linguistic items are lemmas and collocations. Metalinguistic items include the part of speech of a lemma, the paraphrase of its meaning, and syntactic, semantic, and pragmatic indications both on lemmas and on collocations. The dictionary should provide both a lemma and a collocation search as well as a semantic search, using onomasiological categories as metalinguistic patterns that can be chosen and combined by the user. Two kinds of search procedures are available, a simple search and an advanced search procedure. A simple search query should retrieve a single lemma, for instance *paura*, and its entry. Once the search has been performed and the entry has been displayed, the dictionary user can autonomously follow the article-internal hyperlinks to obtain further data. The advanced procedure enables a full-text search on lexicographic articles and on the underlying taxonomic architecture, by means of which lemmas and collocations can be found inside the dictionary.

Placeholder characters can be used for both procedures, either to perform a search that is not targeted at specific results or to obtain multiple results at a time. In the input mask, the search string is made up of lexical content, possibly combined with placeholders. The most powerful placeholder is the wildcard *, which indicates one or more characters, including the null character. It can thus replace both an unknown lexeme in a word combination, i.e., a collocation partner, or a string of characters belonging to a word; in this way the user can find for each lexeme the singular and plural forms, derivative forms, affixed forms,

---

[33] The database was developed by I. Mel'cuk and A. Polguère.

and compounds. Boolean operators are an additional feature of the advanced search procedure. They enhance search flexibility by allowing, excluding, or imposing the presence of specific lexical elements inside a collocation (figure 5).

SEARCH PROCEDURE	INPUT	OUTPUT (excerpt)
simple	paur*	paura, paure
	ans*	ansia, ansie, ansioso, ansiosa, ansiosi, ansiose, ansiogeno, ansiolitico
advanced	ondata di *	ondata di paura, ondata di panico, ondata di terrore, ondata di orrore
	ansia * separazione	ansia di separazione, ansia da separazione
	paura AND buio	paura del buio
	* AND vuoto	paura del vuoto, terrore del vuoto, orrore del vuoto
	panico OR fobia	fobia degli animali, fobia del sangue, panico per il fuoco, panico per il terremoto
	clima NOT ansia	clima di paura, clima di terrore, clima di panico
	film di * NOT terrore	film di paura, film dell'orrore

Figure 5 – Examples of the matches found by employing the wildcard * and Boolean operators

The AND operator states that all of the search words or elements must appear in the results, independent of their syntactic position; the OR operator that one search word or element excludes the presence of the others. From a collocational perspective, AND identifies all collocations that are common to two or more elements, whereas OR identifies all collocations that are not shared by two or more elements. Finally, the NOT operator allows the user to find word co-occurrences that do not contain a specific element. Placeholders and Boolean operators can obviously be combined to obtain more precise search results.

The final output of an advanced search can be further refined by introducing all of the above-mentioned metalinguistic items as search filters. Dictionary data can thus be presented in homogeneous and functional clusters according to the user's specific needs. This is, of course, a key requirement in the case of vast amounts of corpus data. Search filters are syntactic, semantic, or pragmatic, and include various levels of linguistic analysis (figure 6, first three attributes). They originate in the modular system of lemma and collocation description and allow for a corresponding modular representation of search results.

I will now present an example of how an advanced search could be performed with the help of filters. In a text production situation, the dictionary user might need to find all of the collocations containing *paura* or one of its hyponyms in combination with a substantive that indicates a natural element or phenomenon as the origin of that fear. These collocations should be pragmatically unmarked or have a technical use. In the main search mask, the user enters the central item *paura*, whereas all other parameters will be indicated in

specific sub-masks (figure 6[34]). Sub-masks differentiate three fundamental linguistics fields, i.e., syntax, semantics, and pragmatics.

			paura
SYNTACTIC FILTERS	POS	N [V] [A] [ADV]	**N**
	syntactic pattern	**N** + PP (prep. compl.) **N** + AP (quality) **N** (subj.) + V N (subj.) + V + **N** in NP (dir. obj.) N (subj.) + V + **N** in PP (prep. compl.) **N** in PP (prep. compl.) **N** in PP (prep. adj.) **N** in other structures	**N + PP (prep. compl.)**
	[grammatical function]	[subject] [direct object] [prep. compl.] [prep. adj.]	**[no entry possible]**
	[Aktionsart]	[static] [continuative] [telic] [punctual]	**[no entry possible]**
SEMANTIC FILTERS	semantic relation	hypernym hyponym co-hyponym	**AND hyponym**
	thematic role	Experiencer Cause/Agent [Beneficiary]	**Cause/Agent**
	onomasiologic pattern	[animate beings] personified entities abstract entities natural elements/phenomena physical or psychological conditions social/political/economic phenomena [social entities] distressing or dangerous situations/activities bad consequences of one's behaviour	**natural elements/ phenomena**
PRAGMATIC FILTERS	pragmatic marker	none colloquial formal technical figurative literary	**none AND technical**

Figure 6 – Example of an advanced search act with search filters

[34] Bold elements indicate linguistic characteristics of the searched word, not of its collocation partners.

In the first two cases, other subdivisions are required in order to make clear-cut choices possible. Sometimes subfilters are interdependent, sometimes they are not. For example, the syntactic pattern of Class 2, namely NOUN + AP (quality), doesn't require further specification in terms of thematic roles, whereas the pattern of Class 1, namely NOUN + PP (prepositional complement), might. For this reason, during a specific search act, only the filters dependent on a superordinate filter that has been chosen by the user, should be made available for further selection[35].

The POS filter is useful, of course, in the case of homonymic lexemes belonging to different grammatical classes, but in the case of *paura*, which is unequivocally a noun, its value should be automatically entered by the search engine. The POS triggers a set of possible syntactic patterns in which the noun can occur. The selection of the syntactic pattern NOUN + PP (prep. compl.) hinders the choice of a grammatical function and a type of *aktionsart*, which play a role only in the presence of verbal syntagms. In the next step, the user combines the item *hyponym* with the logic operator AND so *paura* won't be rejected from the final results. AND, as we have seen, tags all collocations in which the two coordinated items share at least one collocation partner. The search engine consults the taxonomic hierarchy and retrieves information about the child nodes of *paura*.

The syntactic pattern that has been selected also permits further semantic specification, since the PP can have a different thematic relation to *paura*. Once the user has chosen the option *Cause* and he or she wants to restrict the search to a specific onomasiological type of cause, the next filter can be activated. Finally, pragmatic markers can be identified as well, as long as they fit the user's needs. In the table presented in figure 6, the last attribute on the right collects all of the parameters involved in the search.

The search can now be carried out; its output represents a query-oriented excerpt of the taxonomic and article-internal data contained in the dictionary. The numerous hypertext links allow the user to take a different access path and reach new cross-reference targets at any time, without necessarily performing another search. Figure 7 illustrates the results of the search example.

I endorse the view, expressed by Verlinde/Leroyer/Binon (2010, 4), that in a lexicographic work, there should be a symmetrical focus on data presentation, data access, and user needs. Of course, an unbalanced focus on data access could be the consequence of an overreliance on the electronic medium and of its considerable potential. To avoid the risk, or maybe the temptation, of reducing this electronic dictionary to a mere data search tool, I have tailored the search procedures to the representation of collocations, which in turn depends on the standard demands of the intended user.

An interesting implication of treating collocations as modular syntagmatic constructions with a varying degree of lexical interchangeability is the possibility of combining two or more of them into bigger clusters (cf. Heid 2007, 314–315). The topic of collocation combination has been treated only marginally in the literature, and its importance for understanding how our mental lexicon is built has been substantially underestimated so far. Collocations cannot be considered isolated linguistic atoms: they originate in the most simple syntagmatic levels, but just like single lexemes, they are able to expand by merging with other combinations, creating complex collocational molecules. Bahns (1997, 50–60) points

---

[35] In figure 6, the values in square brackets are not active elements in this specific search.

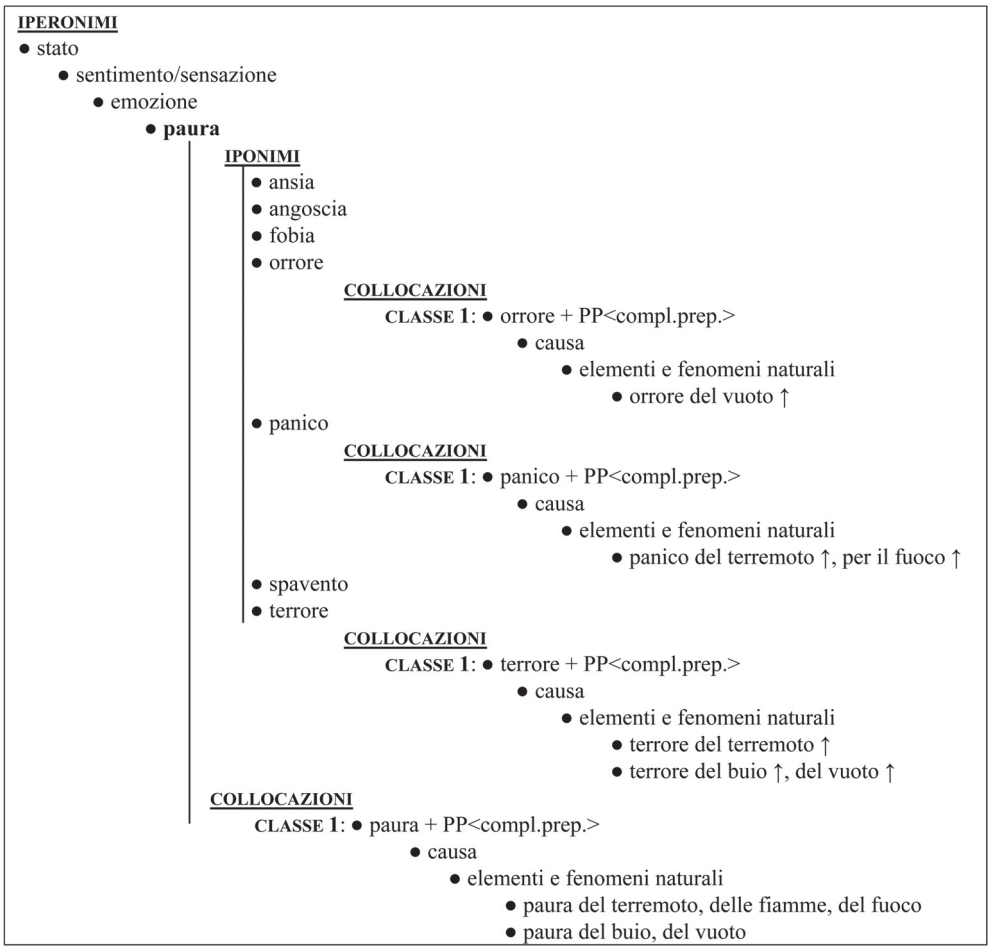

Figure 7 – Search results of the query presented in figure 6

out the interesting possibility of identifying a so-called *Kollokationsfeld*, a collocational field in which collocations can be collected according to different criteria: 1) the merging of collocations with differing syntactic structures and one common collocation partner (*essere preso dalla paura*/*to be seized by fear*, *una paura improvvisa*/*sudden fear*), 2) the presence of collocations with identical structures and one common collocation partner (*paura del vuoto*/*fear of the void*, *terrore del vuoto*/*terror of the void*, *orrore del vuoto*/*horror of the void*), or 3) the presence of collocations with identical structures and different collocational partners with a close semantic affinity (*paura dell'abisso*/*fear of depth*, *terrore del vuoto*/*terror of the void*). The first case concerns exactly the topic of collocation combination mentioned above, whereas the other two present a method for grouping collocations into syntactically and semantically homogeneous sets. Collocations with differing syntactic structures but a common element can be combined as long as the resulting combination observes grammatical restrictions, guarantees pragmatic cohesion, and complies

with general textual requirements such as coherence. In this way, we get the expanded collocation *essere preso da una paura improvvisa* (*to be seized by a sudden fear*), which in fact derives from two smaller collocations. It would be advisable to include this approach to the study of collocations in the search options of an electronic lexicographic tool, allowing the dictionary user to discover how larger syntagmatic structures are often made up of collocational material.

The topic of dictionary search options has been subject to extensive investigation in recent years, and several useful models of search procedures have been developed (cf. among others Nesi 2000, Geeraerts 2000, and de Schryver 2003). In their paper on search techniques in electronic dictionaries, Pastor/Alcina (2010, 319) have the goal of systematising and classifying available search techniques by referring to three elements that are always part of the search process, i.e., the query, the resource, and the result. Their end result is a fine-grained presentation of search techniques, from which we can deduce that search procedures should be flexible enough to be customized according to both to the intended lexicographic function and to the demands of the intended user. As an example, I will mention one important advantage of a corpus-based and database-oriented lexicographic project: with the needs of language professionals such as translators in mind, the search system should be able to retrieve corpus data as contextualised examples of a given linguistic phenomenon. In the dictionary model I propose, corpus co-occurrences are not only the source of dictionary data, they could also serve as ready-to-use lexicographic examples of collocational behaviour. For this reason, every collocation in a dictionary entry should have a hypertext link to the corresponding co-occurrences in the corpus: users would not be provided with only a few examples selected by the lexicographer[36], rather they would have direct access to the corpus data.

6. Conclusions

An electronic dictionary should be a customisable reference tool, specifically tailored to the user's needs. Many authors have attempted to devise typological classifications for electronic dictionaries, tracing the differences between digital and print lexicography. To summarise the observations made in their studies, the most relevant features of an electronic dictionary appear to be multifunctional data access, modular and hypertext data presentation, the customized interface, flexible search capabilities, the closer relationship to contextualised corpus data, interactivity, and multimediality. Existing electronic dictionaries have implemented these features in different ways and using differing nomenclatures, mostly emphasizing one aspect or another. I agree with Pastor/Alcina (2010, 344–345) and their call for the establishment of a more homogeneous classification for evaluating electronic dictionaries, to allow different formats to be easily compared and to improve the teaching of dictionary usage to non-professionals.

---

[36] For a combination of collocations and a restricted set of examples, cf., for instance, the microstructures of DiCE and FrameNet.

In modelling a dictionary of Italian collocations, I have aimed at a flexible representation of collocations: collocational data is intended to serve as an open source inventory, depending strictly on the corpus features. New linguistic and metalinguistic items can be inserted as required. In addition, data classification can easily be adapted to the needs of other user groups, for instance, learners of Italian. Since the lexicographic data come from corpus texts through the mediation of a database, user-oriented modifications mostly take place during macrostructural lemma selection, but also at the level of microstructural and mediostructural implementation. Modularity in the clustering of the data certainly enhances this kind of customisation. Future work should take the form of more such investigations on the interplay of user needs, lexicographic function, and data presentation in electronic dictionaries. With this approach, I hope to contribute to the interesting discussion on adequate lexicographic tools for collocation description, which is far from being concluded.

7. Bibliography

7.1 Dictionaries and lexical databases

BBI = Benson, Morton/Benson, Evelyn, Ilson, Robert, *The Bbi Dictionary of English Word Combinations* (Revised edition), Amsterdam, John Benjamins Publishing Co., 1997.

BLF = Katholieke Universiteit Leuven, *Base lexicale du français*, <http://ilt.kuleuven.be/blf/>.

Cambridge Dictionary of American Idioms = Heacock, Paul (ed.), *Cambridge Dictionary of American Idioms*, Cambridge University Press, 2003.

Cambridge International Dictionary of Idioms = <http://dictionary.cambridge.org>.

DiCE = Facultade de Filoloxía (Universidade da Coruña), *Diccionario de colocaciones del Español*, <http://www.dicesp.com/paginas>.

DiCo = Mel'cuk, Igor/Polguère, Alain, *DiCo, Dictionnaire de Combinatoire*, Observatoire de linguistique Sens-Texte, Université de Montréal, <http://idefix.ling.umontreal.ca/dicofr.html/>, http://olst.ling.umontreal.ca/?page_id=77/lang-pref/en>.

DiCouèbe = Mel'cuk, Igor/Polguère, Alain, *DiCouèbe, Dictionnaire En Ligne De Combinatoire Du Français*, Observatoire de linguistique Sens-Texte, Université de Montréal, <idefix.ling.umontreal.ca/dicouebe>.

FrameNet = University of California (Berkely), *FrameNet*, <http://framenet.icsi.berkeley.edu/>.

Free Dictionary = *The Free Dictionary*, Farlex Inc, <http://idioms.thefreedictionary.com>.

Lexical FreeNet = Datamuse, <http://www.lexfn.com/>.

MultiWordNet = Fondazione Bruno Kessler (FBK), *MultiWordNet*, Human Language Technology Group, Trento. <http://multiwordnet.fbk.eu/english/home.php>.

OCD = *Oxford Collocations Dictionary for Students of English*, Oxford University Press, 2002.

OneLook = <http://www.onelook.com/>.

Ordbog over Faste Vendinger = <http://www.ordbogen.com/opslag.php?dict=fvdd>.

Sabatini-Coletti = *Dizionario Italiano Sabatini-Coletti 2008. Dizionario della lingua italiana*, Milano, Rizzoli Larousse, 2007.

WordNet = Princeton University, *WordNet. A lexical database for English*, <wordnet.princeton.edu>.

Wordsmyth = *Wordsmyth. The Premier Educational Dictionary - Thesaurus*, <wordsmyth.net/>.

Zingarelli = Zingarelli, Nicolo, *Lo Zingarelli 2011. Vocabolario della lingua italiana*, Bologna, Zanichelli, 2010.

7.2 Literature

Bahns, Jens, *Kollokationen und Wortschatzarbeit im Englischunterricht*, Tübingen, Narr, 1997, 39–111.

Baldinger Kurt, *Semasiologie und Onomasiologie*, in: Roland Posner/Klaus Robering/Thomas A. Sebeok (edd.), *Semiotik. Ein Handbuch zu den zeichentheoretischen Grundlagen von Natur und Kultur*, vol. 2, Berlin/New York, de Gruyter, 1998, 2118–2145.

Bartsch, Sabine, *Structural and Functional Properties of Collocations in English. A corpus study of lexical and pragmatic constraints on lexical co-occurrence*, Tübingen, Narr, 2004.

Bergenholtz, Henning, *Das Wortfeld «Angst». Eine lexikographische Untersuchung mit Vorschlägen für ein grosses interdisziplinäres Wörterbuch der deutschen Sprache*, Stuttgart, Klett-Cotta, 1980.

Blumenthal, Peter, *Profil combinatoire des noms. Synonymie distinctive et analyse contrastive*, Zeitschrift für französische Sprache und Literatur 112 (2002), 115–138.

Blumenthal, Peter, *Wortprofil im Französischen*, Beihefte zur Zeitschrift für romanische Philologie 332 (2006).

Blumenthal, Peter, *Französische Kollokationen in Lexikografie und Forschung*, Lexicographica 24 (2008), 21–38.

Conway, Martin A./Bekerian, Debra A., *Situational Knowledge and Emotions*, Cognition and Emotion 1/2 (1986), 145–191.

Cowie, Anthony Paul, *Introduction to the «Oxford Dictionary Of Current Idiomatic English»*, in: id., *Oxford Dictionary Of Current Idiomatic English. Phrase, Clause & Sentence Idioms*, Oxford, Oxford University Press, 1983, XII-XIII.

Cowie, Anthony P. (ed.), *Phraseology. Theory, Analysis, and Applications*, Oxford, Clarendon Press, 1998, 1–53.

de Schryver, Gilles-Maurice, *Lexicographers' Dreams in the Electronic-Dictionary Age*, International Journal of Lexicography 16/2 (2003), 143–199.

Ekman, Paul, *An Argument for Basic Emotions*, Cognition and Emotion 6 (3/4) (1992), 169–200.

Fillmore, Charles J., *Schemata and Prototypes. Lecture notes of a symposium held at Trier University, 1977*, in: Rene Dirven/Günter Radden (edd.), *Fillmore's Case Grammar. A Reader*, Heidelberg, Groos, 1977, 99–106.

Geeraerts, Dirk, *Adding Electronic Value. The Electronic Version of the Grote Van Dale*, in: Ulrich Heid (edd.), *Proceedings of the Ninth Euralex International Congress, EURALEX 2000*, vol. 1, Stuttgart, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 75–84.

Hausmann, Franz Josef, *Le dictionnaire de collocations*, in: Franz Josef, Hausmann, et al. (edd.), *Wörterbücher. Ein internationales Handbuch zur Lexikographie*, vol. 1., de Gruyter, 1989, 1010–1019.

Hausmann, Franz Josef, *Was sind eigentlich Kollokationen?*, in: Kathrin Steyer (ed.), *Wortverbindungen - mehr oder weniger fest*, Berlin/New York, de Gruyter, 2004, 309–334.

Heid, Ulrich et al., *Struktur und Interoperabilität lexikalischer Ressourcen am Beispiel eines elektronischen Kollokationswörterbuchs*, in: Lothar Lemnitzer/Georg Rehm/Andreas Witt (edd.), *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen*, Tübingen, Narr, 2007, 313–322.

Heringer, Hans Jürgen, *Das höchste der Gefühle: empirische Studien zur distributiven Semantik*, Tübingen, Stauffenburg, 1999.

Johnson-Iaird, Philip N./Oatley, Keith, *The Language of Emotions. An Analysis of a Semantic Field*, Cognition & Emotion 3/2 (1989), 81–123.

Kammerer, Matthias, *Hypertextualisierung gedruckter Wörterbuchtexte. Verweisstrukturen und Hyperlinks. Eine Analyse anhand des «Frühneuhochdeutschen Wörterbuchs»*, in: Angelika Storrer/Bettina Harriehausen (edd.), *Hypermedia für Lexikon und Grammatik*, Tübingen, Narr, 1998, 144–171.

Lyons, John, *Semantics*, Cambridge, Cambridge University Press, 1977.

Müller-Spitzer, Carolin, *Vernetzungsstrukturen lexikografischer Daten und ihre XML-basierte Modellierung*, Hermes 38 (2007), 137–171.

Nesi, Hilary, *Electronic Dictionaries in Second Language Vocabulary Comprehension and Acquisition: the state of the art*, in: Ulrich Heid, et al. (edd.), *Proceedings of the Ninth Euralex International Congress, EURALEX 2000*, vol. 2, Stuttgart, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 839–847.

Pastor, Verónica/Alcina, Amparo, *Search Techniques in Electronic Dictionaries: A Classification for Translators*, International Journal of Lexicography 23/3 (2010), 307–354.

Reichmann, Oskar, *Das onomasiologische Wörterbuch: Ein Überblick*, in: Franz Josef Hausmann, et al. (edd.), *Wörterbücher. Ein internationales Handbuch zur Lexikographie*, vol. 2., de Gruyter, 1991, 1057–1067.

Rovere, Giovanni, *Correspondances et équivalences: fr. «orgueil/superbe» – it. «orgoglio/superbia»*, in: Peter Blumenthal/Salah Mejri (edd.), *Les séquences figées: entre langue et discours*, Steiner, Stuttgart, 2008, 159–174.

Schafroth, Elmar, *Kollokationen im GWDS*, in: Herbert Ernst Wiegand (ed.), *Untersuchungen zur kommerziellen Lexikographie der deutschen Gegenwartssprache I. «Duden. Das große Wörterbuch der deutschen Sprache in zehn Bänden»*, Tübingen, Niemeyer, 2003, 397–412.

Strapparava, Carlo/Valitutti, Alessandro/Stock, Oliviero, *The Affective Weight of Lexicon*, in: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 2006, 1–83.

Tarp, Sven, *Lexicography in the borderland between knowledge and non-knowledge: general lexicographical theory with particular focus on learner's lexicography*, Tübingen, Niemeyer, 2008.

Tissari, Heli, *On the concept of sadness: looking at words in contexts derived from corpora*, in: Barbara Lewandowska-Tomaszczyk (ed.), *Corpus Linguistics, Computer Tools and Applications: State of the Art*, Frankfurt am Main, Lang, 2007, 291–308.

Tognini-Bonelli, Elena, *Corpus Linguistics at Work*, Amsterdam, Benjamins, 2001.

Valitutti, Alessandro/Strapparava, Carlo/Stock, Oliviero, *Developing Affective Lexical Resources*, Psychology 2/1 (2004), 61–83.

Verlinde, Serge/Leroyer, Patrick/Binon, Jean, *Search and You Will Find. From Stand-Alone Lexicographic Tools to User Driven Task and Problem-oriented Multifunctional Leximats*, International Journal of Lexicography 23/1 (2010), 1–17.

Wiegand, Herbert Ernst, *Synonyms Appearing in Major Alphabetical Dictionaries of Contemporary German*, in: Antje Immken/Werner Wolski (edd.), *Semantics and Lexicography. Selected Studies (1976–1996)*, Tübingen, Niemeyer, 1999, 139–151 (= 1999a).

Wiegand, Herbert Ernst, *Are Most of the Cumulative Dictionaries of Synonyms really Useless Bestsellers?*, in: Antje Immken/Werner Wolski (edd.), *Semantics and Lexicography. Selected Studies (1976–1996)*, Tübingen, Niemeyer, 1999, 283–296 (= 1999b).

Wiegand, Herbert Ernst, *Altes und Neues zur Mediostruktur in Printwörterbüchern*, Lexicographica 18 (2002), 168–252.

Wiegand, Herbert Ernst, *Wörterbuchbenutzung bei der Übersetzung. Möglichkeit ihrer Erforschung*, in: Vida Jesenšek (ed.): *Wörterbuch und Übersetzung: 4. Internationales Kolloquium zur Lexikographie und Wörterbuchforschung (Universität Maribor 20. bis 22. Oktober 2006)*, Hildesheim/Zürich/New York, Olms, 2008), 1–43.

Wierzbicka, Anna: *«apples» are not a «kind of fruit»: the semantics of human categorization*, American Ethnologist 11/2 (1984), 313–328.

Wierzbicka, Anna, *Angst*, Culture and Psychology 4/2 (1998), 161–188.