A proposal for an electronic dictionary of Italian collocations highlighting lexical prototypicality and the syntactic-semantic relations between collocation partners

Laura Giacomini
Seminar für Übersetzen und Dolmetschen (SÜD) am Institut für Allgemeine u. Angewandte Sprach- und Kulturwissenschaft (IASK), Ruprecht-Karls-Universität Heidelberg

This paper presents a corpus-based case study aimed at designing an electronic dictionary of Italian collocations and focussing on a small set of nouns belonging to the semantic field of paura/ fear. In the paper, all of the steps involved in data retrieval, automatic and non-automatic evaluation, collocation selection and lexicographic organization are explained in detail. Lexicographic data are represented as a three-dimensional lexical framework displaying ontological, semantic and syntactic relations among lexemes. On the ontological level, paura as an entity is connected to contiguous emotions, but it also serves as a prototype for the category of the lexemes selected, shaping their syntactic and semantic behaviour. Collocations are formally categorized through a set of analytic parameters which enable a detailed lexical description as well as more finely grained dictionary search results. On the microstructural level, substantival collocation partners of the selected nouns are described in terms of thematic roles and semantic features, whereas adjectival collocation partners additionally have a set of principles derived from psychological studies applied to them. Finally, analysis of verbal collocation partners focusses on the interplay of the grammatical function and the thematic role of the noun they cooccur with, and on verbal Aktionsart.
The intended lexicographic description rests upon a coherent cross-reference network linking the prototype to the other lemmas, collocation partners to each other, and collocations belonging to the narrower semantic field of paura, as well as collocations belonging to other semantic fields (e.g., the semantic field of emotions). At the same time, according to an open source principle, corpus-based lexical data can be deductively expanded using framework information.

1. Introduction

The aim of this case study is to design an electronic dictionary of Italian collocations addressed to professional translators and advanced learners[1]. In the Italian lexicographic landscape, a dictionary of collocations remains a desideratum. The representations of collocations in the general language dictionaries currently available on the market are largely defective or at least unclear. Such dictionaries primarily lack syntactic information as well as a clear-cut distinction between free lexical combinations and proper collocations. Lemmatisation principles applying to multiword expressions are not always comprehensible and rigorous, and the semantic relations existing among different collocations are entirely neglected. Whenever a combination is missing, for whatever reason, users are unable to predict that such a combination is likely to be a collocation. But this is also true of dictionaries of collocations already existing for other languages. Dictionary users should therefore be provided with a new, specific resource for textual production and language skills improvement.

2. Description of the case study

Dictionary data were retrieved from a non-annotated electronic reference corpus of major Italian newspapers totaling about 300 million words (tokens), and are intended to serve as a corpus-based open inventory, which could be extended by the user according to specific syntactic, semantic and pragmatic rules. A lexicographic corpus, made up of the most common general language Italian dictionaries and some dictionaries and encyclopedias of psychology and philosophy, was also available for comparison purposes.

[1] This research is part of my PhD project at the University of Heidelberg's Institute for Translation and Interpreting.

The case study focuses on a set of nouns belonging to the semantic field *paura*[2] (*fear*): *paura* (*fear/worry*), *ansia*/*angoscia* (*anxiety*), *panico* (*panic*), *fobia* (*phobia*), *spavento* (*fear*/*fright*), *terrore* (*terror*), *orrore* (*horror*/*scare*). Lexicographic items are represented within a three-dimensional lexical network displaying ontological (onomasiological), semantic and syntactic relations among lemmata and among collocations. Online resources such as FrameNet and WordNet served as a model for lexical annotation.

2.1. A functional definition of collocation

An essential requirement for the representation of the selected nouns was the choice of a functional definition of collocation. In a narrow sense, collocations were defined as idiomatic multiword expressions (*niente paura*/*don't worry*, *paura matta*/*jitter*, *paura mortale*/*mortal fright*), subject to restricted compositionality, substitutability and modifiability, and identifiable through standard idiom tests. In a broad sense, I considered collocations as familiar word combinations recurring in our mental lexicon (*paura del buio*/*fear of the dark*, *paura di volare*/*fear of flying*), mostly associated with typical scenes, i.e., situational contexts in Fillmore's sense (Fillmore 1977). Provided that corpus data are initially assigned a specific collocational status on the basis of statistical significance and subsequent qualitative selection, strict boundaries between different degrees of phraseological cohesion are no longer required. Moreover I have used the notion of *phraseologism* as proposed by Burger (2007), meaning a word combination which is characterized by polylexicality and fixedness; a third feature, idiomaticity, comes into play when phraseology is intended in a narrow sense.

Analytical approaches to the study of collocation were abandoned in favour of a more lexicographic-oriented view. As opposed to Hausmann's conception, for example, I have not limited myself to considering binary word combinations (Hausmann 1999). Furthermore, despite the clear identification of a semantically autonomous collocation constituent alongside one or more subordinate constituents, the concepts of *base* and *collocate* were rejected; for one example of the reasoning behind this choice, they didn't prove useful when considering minimal pairs such as *ansia del momento* (*current*/*momentary anxiety*) and *momento d'ansia* (*moment of anxiety*).

3. Methods of data evaluation

3.1. Statistical evaluation and manual procedure

Statistical significance of the extracted co-occurrences was measured by means of the log-likelihood function: by comparing the hypotheses of independence and dependence for single lexical items, the likelihood ratio determines the extent to which word combinations are expected to occur in the corpus. The likelihood ratio test proved to be a useful means to detect sparse data and to reject many semantically general words. Part-of-speech tags were then added to the statistically relevant data, producing better clustering results.

I evaluated the extracted collocation candidates through a stepwise manual procedure applied to automatically retrieved information such as likelihood ratios, frequency values, syntactical (positional) features and POS tags. Manual evaluation enables the identification of particular phenomena like, for instance, corpus-specific properties (high likelihood or frequency values for words accidentally co-occurring with *paura* and other nouns), or significant likelihood values for words which are not proper collocation partners of a given noun but rather belong to one of its (stereo)typical scenes (cf. *paura/fear* + *incendio*/*fire*, *fumo*/*smoke*). The main

[2] Hereafter are indicated the most common English equivalents for each Italian lexeme. These equivalents are obviously not meant to provide the reader with an exhaustive list of translation possibilities.

advantage of this method is the ability to detect lexicographically relevant phenomena such as complex predicates (*avere paura*) and non-binary collocations (*prendersi un bello spavento*), which the bigram-based likelihood method cannot account for.

3.2. A brief description of the steps involved in non-automatic evaluation

The first step of the evaluation consisted of sorting out 'naively' irrelevant occurrences such as spelling mistakes, newspaper section names, lexemes with no logical or syntactical link to the selected nouns as well as idiolectal and intertextual utterances, or words exclusively related to particular news and events. Then I surveyed interference effects between collocations, in particular the exchange of collocational partners: *gettare nella paura*/*nel panico* (*to send into panic*), *la paura*/*la rabbia esplode* (*rage explodes*). Finally I assessed the degree of lexicalisation of collocation candidates by testing their morphosyntactic fixedness through conventional substitutions (pronominalisation, anaphora, interrogative or relative clause) and modifications (adjectival modifier, plural form). In this way candidates become eligible for dictionary recording and can be further prepared to conform to lexicographic representation. Word combinations available in the lexicographic corpus were integrated into the data set extracted from the newspaper corpus. In order to avoid arbitrary choices, no other collocations were added to the ones found in the corpora, although the inclusion of many lexical combinations would have been justified at least by intuition.

4. Prototypicality and the three-dimensional structure

In the above-mentioned three-dimensional structure, the ontological level takes into account the process of classification through which we establish whether two or more items (physical objects, properties, actions) are members of the same category. According to the traditional view of classification, category items are clearly marked by necessary and sufficient conditions, while category boundaries are rigidly drawn. Nevertheless, many concepts (or mental representations of category items) are not easily definable, and their features are only probable. In this study, I will therefore apply the prototype view as developed in philosophy and psychology in the early 1970s as an alternative to the classical approach (cf. Rosch 1977). In the field of cognitive emotion studies, prototypes have already been applied to basic human emotions, i.e., a primitive set of emotions, including fear, anger, happiness, sadness or disgust, which have universal bodily or phenomenological components such as specific facial expressions (cf. Johnson-Laird and Oatley 1989, Ekman 1992). At the heart of the prototype theory is the idea that category membership is determined by sufficient resemblance to prototypical items, i.e., the *best examples* of a given class. Categories are not represented in terms of defining properties but as fuzzy sets of elements. Prototypicality works well for both extralinguistic and linguistic categorization, allowing for clearer cross-referencing between a semantically central lexical item (*paura*: the prototype) and the elements belonging to its category, namely, the other substantives.

This approach makes up for missing or incomplete lemmatic and collocational information in existing lexicographic publications. The ontological dimension regulates, on an external, macrostructural level, the relations between semantically contiguous entities and concepts (*paura* is onomasiologically linked to other basic emotions and to secondary, i.e., complex, emotions). This applies to the dictionary-making process in a wider perspective.

5. Pre-dictionary analysis: syntactic and semantic categorization of collocation partners

On a microstructural level, the connection between the syntactic and semantic dimensions is exemplified by the fact that differences in word and collocation meanings often strictly depend on differences in syntagmatic structures. For this reason, it was necessary to provide a detailed description of the various syntagmatic frames in which *paura* and the other nouns may occur, which also allowed for an initial comparison of the selected nouns and pointed out the prototypical role of *paura* in shaping the syntactic behaviour of the lexical group.

According to the Fregean view, *paura* and the other nouns correspond to unsaturated entities and have no complete, independent meaning. Their meaning has to be contextually supported by collocation partners. Analytic parameters have been introduced as a notational method for the purpose of specifying the semantic features of collocation partners and of classifying these into homogeneous groups. Parameter typology is closely linked to syntactic (part-of-speech) functions.

5.1. Substantives and adjectives

Substantival collocation partners are best described in terms of thematic relations and semantic features. A simplified set of thematic relations includes the major distinctive items <Agent/Cause>, <Experiencer>, <Patient/Theme> and <Beneficiary>. Since no clear boundaries can be traced between relations, further specification might produce unfavourable effects on lexical description. Semantic features complete thematic information in the form of mutually exclusive labels assigned to entities: [entity: ± concrete], [+ concrete: ± animated], [+ animated: ± human].

Adjectival collocation partners can be identified both through thematic relations and by a framework of principles and constraints derived from psychological studies: they concern the origin, nature, intensity, duration and adequacy of emotions. Aktionsart serves as a further parameter for specifying the semantic features of verbal collocation partners.

Table 1 illustrates parameter types applied to substantival (N) and adjectival (A) collocation partners of *paura*, and contains some collocation examples from the corpus.

If we consider, for instance, *paura* as requiring a prepositional phrase or a clause (cf. Table 1, second and third column), analytic parameters can be applied to distinguish the semantic function of the post-head string in the PP. On the one hand, the head *di* (*of*) can take a noun (*paura del vuoto*/*fear of void*) or a verb in the infinitive form (*paura di volare*/*fear of flying*) indicating emotion-triggering entities or events which can be labeled as <Cause> [± concrete]. On the other hand, *di* can combine with a noun (*paure dei genitori*/*parental fears*) referring to the experiencing entity: <Experiencer> [+ animated]. The alternative head of the PP *per* (*of*/*for*) can introduce either a <Cause> [- concrete]/[- animated] (*paura per il risultato*/*fear of the result*) or a <Beneficiary> [+ concrete] (*paura per i figli*/*concern for the children*).

		paura head of a NP with A-modifier	*paura* + PP	*paura* + Clause
<Agent/Cause>	A/N [+ animated] human beings/animals (groups)	p. islamica	p. dei ladri	
	A self-/reality perception N [- concrete][- animated] abstract entities, events N [- concrete][- human] natural phaenomena	p. esistenziale	p. per il risultato p. del vuoto	p. di volare
	A/N [- concrete] historical/evolutionary origin	p. ancestrale	∅	
	A/N [- concrete] pathological origin	p. fobica	∅	
	A/N [- concrete][- animated] social/political/economic origin	p. razziale		
<Experiencer>	A/N [+ animated] human beings/animals (groups)	p. collettiva	p. dei genitori	
	A/N [- concrete][- animated] personification of a social/political/economic aspect	∅	p. del governo	
<Beneficiary>	A/N [+ concrete]	∅	p. per i figli	
<∅>	A adequacy of emotion	p. infondata		
	A intensity/duration of emotion	p. profonda		

Table 1. Substantival/Adjectival collocation partners of *paura* (excerpt from the corpus)

Paura can also be the head of a nominal phrase (NP) including adjectival modifiers (cf. Table 1, first column) indicating a <Cause>, i.e., the origin/nature of the emotion (*paura fobica*/*phobic fear*), an <Experiencer> (*paura collettiva*/*collective fear*), the intensity or duration of the emotion (*paura profonda*/*deep fear*), or its level of adequacy as perceived by observers (*paura infondata*/*groundless fear*). Further distinctions have been introduced into these predefined classes in order to obtain a more fine-grained adjectival description: for instance, a <Cause> relation can be understood in terms of self- or reality perception (*esistenziale*/*existential*), historical/evolutionary origin (*ancestrale*/*ancestral*), social/political/economic origin (*razziale*/*racial*), or pathological origin (*ossessiva*/*obsessive*).

Table 2 shows, in part, the analysis parameters used for the description of verbal collocation partners, with some real collocation examples for *paura*.

	paura as:		Aktionsart:
	subject	p. cresce	telic
		p. dilaga p. serpeggia	continuative
subj. other than *paura*:			
<Agent/Cause>	direct object	fare p. mettere p.	stative/continuative
		generare p.	telic
	prepositional complement	fare leva sulle p.	stative/continuative
<Experiencer>	direct object	avere p. provare p. avvertire p.	stative/continuative
	prepositional complement	morire di p.	telic/punctual
		tremare di p.	continuative

Table 2. Verbal collocation partners of *paura* (excerpt from the corpus)

5.2. Verbs

Analysis of verbal collocation partners (cf. Table 2) focuses on the interplay of a) the grammatical function of a given noun, for instance *paura*, as a subject, direct object or prepositional complement, b) the thematic relation assigned to subjects other than *paura*, and c) verbal Aktionsart. If *paura* does not take the subject role, it may function as an object or a complement and combine with a simple verbal lexeme, forming a complex predicate. Complex predicates whose subjects introduce <Experiencer> and <Agent/Cause> relations mostly express the scripts, i.e., the semantic core, behind unsaturated nouns like *paura*:

(1) *qu* <Experiencer> *ha paura di qu/qc* <Agent/Cause> (*sb fears sb/sth*) and
(2) *qu/qc* <Agent/Cause> *fa paura a qu* <Experiencer> (*sb/sth scares sb*).

Not only do compound verbs such as *avere/provare/avvertire paura* share structure (1), they also tend to share Aktionsart features. They are inherently durative (atelic) and express either a state or an activity. Depending on the kind of action involved, verbal aspect can be imperfective (*aveva paura*) or perfective (*ebbe paura*). Finally, pragmatic markers specifying register, style or terminological information have been added to syntactic and semantic tags to complete the lexical description.

5.3. Other significant combinations

Following the common distinction made between lexical collocations and grammatical collocations, cooccurrences of one substantive and a preposition (*per paura/out of fear*, *senza paura/without fear*, *fearless*) or an exclamatory adjective (*che spavento!/what a fright!*) were analysed and listed separately.

Idioms were also grouped in separate clusters. Idiomatic combinations are often non-binary expressions with a complex syntagmatic structure including different parts of speech. From a morphosyntactic and semantic perspective, they are relatively frozen units whose literal meaning and phraseological meaning tend to diverge to some degree: cf. for example the difference between *morire di paura* (which has both a literal and a phraseological reading: *to*

die out of fear/to be scared stiff, to be terror-stricken) and *la paura fa novanta* (which has only one phraseological reading: *fear is present, fear is all around*).

6. The microstructural perspective

After selecting and organizing the extracted data according to coherent syntactic and semantic principles, the next step concerns the choice of appropriate structures for presenting data in the form of dictionary entries.

On the abstract microstructural level, a dictionary accounts for the following information:

1) The reference word (i.e., the prototypical lexeme, *paura*) functions as the nodal point between prototypes belonging to other ontological entities (for instance, other basic emotions such as anger, happiness or disgust together with the corresponding lexical items) and the elements of its category (*ansia*, *angoscia*, *panico*, *fobia*, *terrore*, *orrore*, *spavento*). Dictionary data are organized around the prototype itself. The process of lemmatisation involves the selected nouns, whereas collocation partners are coinceived as reference related links and not as independent lemmas.

2) Lexicographic description focuses on collocations as semantic units; the traditional definition for a lemma is therefore substituted by a complex set of information concerning each collocation as a whole. Starting from the prototype *paura*, all collocations are represented in terms of part of speech, syntactic and semantic parameters as described in § 5, pragmatic labels and degree of idiomaticity. Additionally, each collocation partner refers the dictionary user to all other substantives with which it forms a collocation. In Table 3, each down arrow indicates a cross-reference to similar collocations: for instance, *paura generale* is linked to *panico generale* and *spavento generale*. Positional features are also pointed out whenever possibile: for example, an A is likely to form a marked or a not marked combination with a substantive, depending on their relative position.

PAURA		**A + *paura***	***paura* + A**	***paura* + PP/Clause**	
<Exp>	A/N [+ animated] human beings/animals (groups)		generale ►[rabbia] ▼ [panico] [spavento] collettiva ▼ [panico] metropolitana ►[leggenda]	p. della gente p. della popolazione	p. tra la popolazione ▼ [terrore] [timore]

Table 3. Excerpt from the abstract microstructure

3) *Paura* is the only lemma with an autonomous entry. Collocation partners should be referenced to the prototypical lexeme. In the above mentioned example, *panico generale* implies a double reference goal, namely, *panico* > *paura* and *panico generale* > *paura generale*. At the same time, every possible connection to other sets of lexemes should be pointed out (cf. right arrows in Table 3): *paura generale* > *rabbia generale*, *paura*

metropolitana > *leggenda metropolitana*. Obviously, this is beyond the scope of my case study, because it involves different semantic fields.

The concrete lexicographic microstructure accounts for substantial differences in the representation of dictionary search results. The most important advantage of an electronic dictionary of collocations lies in the easy retrieval of selective information. On the one hand, dictionary users can search for a substantival lemma and receive as output all of its collocations, or a set of collocations according to specific search parameters. On the other hand, every collocation partner (*paura*, *panico*, *generale*, *metropolitano*, *gente*, …) as well as every multiword expression (*paura collettiva,…*) can be looked up separately.

<table>
<tr><td colspan="2">QUERY [result: not found!]

le ansie (subject) serpeggiano (continuative V)</td></tr>
<tr><td>ansie + other V:</td><td>serpeggiano + other subjects:</td></tr>
<tr><td>continuative V:
-</td><td rowspan="2">la paura serpeggia
le paure serpeggiano
l’ansia serpeggia
le angosce serpeggiano
il panico serpeggia
il terrore serpeggia</td></tr>
<tr><td>telic V:
le ansie assalgono ► [la rabbia assale qu]
[la gioia assale qu]
▼
[la paura assale qu, qu è assalito da paura]
[le paure assalgono qu]
[l’ansia assale qu]
[il panico assale qu]
[qu è assalito dallo spavento]
[il terrore assale qu]</td></tr>
</table>

Table 4. Query result for *le ansie serpeggiano/fears go around* (excerpt)

As shown in Table 4, even if a collocation is missing from the corpus (for instance, *le ansie serpeggiano*), dictionary users are nevertheless presented with syntactically equivalent structures (*ansie* as a subject of other predicates) and semantically interchangeable paradigms (other substantives serving as subjects of *serpeggiare*) from among which they can choose, or from which they can predict whether a certain cooccurence might be classified as a collocation.

7. Conclusions

This procedure aims at a formal categorization of collocation partners, allowing for more finely grained dictionary search results. Dictionary users can recognise a lexical item as being part of a certain collocational framework. At the same time, according to an open source principle, corpus-based lexical data can be deductively expanded by using framework information. The intended lexicographic description rests upon a coherent cross-reference network, linking

(i) the prototype to the other lemmata,
(ii) collocation partners,
(iii) collocations belonging to the narrower semantic field of *paura*,
(iv) collocations belonging to a broader semantic field (e.g., semantic field of emotions).

The electronic medium opens up the possibility of a complex cross-reference system. Nevertheless, lexical information which meets search requirements is made available to dictionary users in the form of simple excerpts from the concrete microstructure.

References

Burger, H. (2007). *Phraseologie: eine Einführung am Beispiel des Deutschen*. 3. Auflage. Berlin: E. Schmidt.

Ekman, P. (1992). 'An Argument for Basic Emotions'. In *Cognition and Emotion* 6 (3/4). 69-200.

Fillmore, C.J. (1977). 'Scenes-and-frames semantics'. In: Zampolli, A. (ed.). *Linguistic Structures Processing*. Amsterdam/New York/Oxford: North-Holland Publishing Company. 55-81.

Frege, G. (2007). *Funktion, Begriff, Bedeutung*. Hrsg. von Mark Textor. 2. Auflage. Göttingen: Vandenhoeck-Ruprecht.

Hausmann, F.J. (1999). 'Praktische Einführung in den Gebrauch des Student's Dictionary of Collocations'. In: Benson, M. et al. *Student's Dictionary of Collocations*, Berlin: Cornelsen. iv-xiii.

Johnson-Laird, P.N.; Oatley, K. (1989). 'The Language of Emotions. An Analysis of a Semantic Field'. In *Cognition and Emotion* 3 (2). 81-123.

Rosch, E.H. (1977). 'Human Categorization'. In Warren, N. (ed.). *Studies in Cross-cultural Psychology*. Vol 1. London/New York/San Francisco. 1-49.