Towards an Empirical Subcategorization of Multiword Expressions

Luigi Squillante
Dipartimento di Scienze Documentarie, Linguistico-filologiche e Geografiche
“Sapienza” - Università di Roma
Roma, Italy
luigi.squillante@uniroma1.it

Abstract

The subcategorization of multiword expressions (MWEs) is still problematic because of the great variability of their phenomenology. This article presents an attempt to categorize Italian nominal MWEs on the basis of their syntactic and semantic behaviour by considering features that can be tested on corpora. Our analysis shows how these features can lead to a differentiation of the expressions in two groups which correspond to the intuitive notions of multiword units and lexical collocations.

1 Introduction

In contemporary linguistics the definition of those entities which are referred to as multiword expressions (MWEs) remain controversial. It is intuitively clear that some words, when appearing together, have some “special bond” in terms of meaning (e.g. black hole, mountain chain), or lexical choice (e.g. strong tea, to fill a form), contrary to free combinations. Nevertheless, the great variety of features and anomalous behaviours that these expressions exhibit makes it difficult to organize them into categories and gave rise to a great amount of different and sometimes overlapping terminology.1 In fact, MWEs can show non-grammatical constructions, syntactic fixedness, morphological frozeness, semantic restrictions, non-compositionality, strong pragmatic connotation, etc. These features are not necessary and sufficient conditions for each expression, but represent only possible behaviours that can be exhibited together or individually and to a different extent.

Traditionally MWEs are seen as entities lying on a continuum between two poles that go from a maximum of semantic opacity (green thumb) to compositional expressions that show only lexical restrictions (to catch a cold). However the “compositional criterion” is a problematic concept in semantics, since it has been shown how difficult it is, in language, to define component parts, rules or functions involved in compositionality (Casadei, 1996) and, above all, that it is impossible to give words an absolute meaning independently from their context (Firth, 1957; Hanks, 2013). Because of this, the problem of subcategorizing the heterogeneous set of MWEs must be based on more reliable and testable criteria.

This work presents a study conducted on the Italian language that aims at dividing MWEs in subcategories on the basis of empirical syntactic and semantic criteria different from compositionality. We show how these features are able to separate two poles of entities which approximately correspond to what is intuitively known as multiword units (polirematiche in the Italian lexicographic tradition)2 as opposed to (lexical) collocations.

2 The need to go beyond statistics

In recent years, the fact that MWE components tend to cooccur more frequently than expected led to the development of several statistical association measures3 (AMs) in order to identify and automatically extract MWEs. However, as pointed out in Evert (2008), it is important not to confuse the empirical concept of recurrent or statistically relevant word combination in a corpus (empirical collocation) with the theoretical concept of MWE (which assumes phraseological implications), although the two sets overlap. In fact, it is common

---

1See Bartsch (2004) or Masini (2007) for an overview on the historical development of MWE terminology.
2cf. De Mauro (2007).
3See Evert (2004) for a general overview.
that AMs can extract expressions such as leggere un libro ‘to read a book’ or storcere il naso ‘to stick up [one’s] nose’ just because the components tend to cooccur often in corpora. However, while the first one seems not to need its own categori
cal status (Bosque, 2004), the latter is usually de
noted as a metaphoric MWE or idiom. AMs are
not able to distinguish between the two or even
differentiate subtypes of true MWEs on the basis
of phraseological relevance (e.g. AMs are not able
to assign a higher score to more opaque MWEs
in opposition to lexical collocations). It is pos
sible, however, to integrate statistical information
with the results of syntactic and semantic tests per
formed on corpora in order to identify subgroups
of MWEs.4

3 Methodology

As a first approach, in this work only Italian nomi
nal MWE of the form [noun + adjective]5 are cho
sen. The corpus used in our study is PAISÀ6, a
freely available large Italian corpus, composed of
ca. 250 million tokens and morpho-syntactically
notated. By means of mwetoolkit (Ramisch et
al., 2010) the 400 most frequent [noun + adjecti
ve] bigrams are extracted from the corpus and
assigned the pointwise mutual information (PMI)
association score (Church and Hanks, 1990). Then
the bigrams are ordered according to PMI and only
the first 300 are retained.7 The number of oc
currences of the expressions contained in this set
varies between 20,748 and 641.

Then, we implemented a computational tool
that performs empirical tests on modifiability. We
chose to study three features, which are a) interrup
tibility, b) inflection and c) substitutability8 and
for each of them an index is calculated.

Given the expression, the index of interruptibil
ity (Ii) compares the occurrences of the sequence
in its basic form [noun + adjective] (nbf), with
the occurrences of the same sequence with one
word occurring between the two components (ns).
The queries are made over lemmas and its value is
given by the ratio: \[ I_i = \frac{n_i}{n_{bf} + n_i} \].

The index of inflection (If) compares the num
ber of occurrences of the prevalent (most frequent)
inflected form (n pf) with those of the basic lemm
atized form (nbf) and its value is given by the
ratio: \[ I_f = \frac{n_{bf} - n_{pf}}{n_{bf}} \].

Finally, the index of substitutability (Is) com
pares the number of occurrences of the basic form
(nbf), regardless of inflection, with the occurr
cences ns of all the sequences in which one of
the two components is replaced by one of its
synonyms (if present). If ns,i is the number of oc
currences of the i-th synonym of the first com
ponent word and ns,2,i is an analogous quan
tity for the second component word, then \[ n_s = \sum_i n_{s,1,i} + \sum_i n_{s,2,i} \] and \[ I_s = \frac{n_s}{n_{bf} + n_s} \]. In
order to calculate Is, the tool needs an external syn
onym list; we chose the GNU-OpenOffice Italian
Thesaurus10 because of its immediate availability,
open-source nature and ease of management.11

Then the three indices are calculated for each of
the 300 MWEs of the candidate list.

4 Results

Figure 1 shows the distribution of the expres
sions in the planes defined by Ii, If, Is. It is eviden
t that there is a tendency for the expres
sions to gather more along the axes rather than in the planes, i.e.
where one of the indices has low values.

For the insertion of other words between the com
ponents (e.g. carro armato ‘tank’; cf. *carro grande armato) while others
do (e.g. punto debole ‘weak point’; cf. punto più debole); b) some nominal MWEs exhibit inflection frozeness (e.g. diritti umani ‘human rights’; cf. *diritto umano), while others can be freely inflected (e.g. cartone animato ‘cartoon’; cf. cart
toni animati); c) some nominal MWEs do not allow for the
substitution of one of their components with a synonym (e.g.
colonna sonora ‘soundtrack’; cf. *pilastro sonoro) while oth
ers do (e.g. guerra mondiale ‘world war’; cf. conflitto mondiale).

Although Nissim and Zaninello (2011) show how Ital
ian nominal MWEs can exhibit several distinct morphologi
cal variations, we chose to consider only the proportion be
tween the prevalent form and the total number of expres
sions since our pattern generally admits only singular and plural
forms, with noun and adjective coherently coupled.

10http://linguistico.sourceforge.net/pages/thesaurus_italiano.html

11However, other more specific and complete resources
could be attached instead in the future, in order to improve
the quality of the results.
Since the plane \( I_f I_s \) shows the highest dispersion of points, we chose to consider in this plane 4 polarities defined by the intersection of high/low values for both \( I_f \) and \( I_s \). We consider a value high (and indicate \( I^+ \)) when \( I > 0.33 \) and low (\( I^- \)) when \( I < 0.1 \). In this way we obtain 4 sets of expressions lying at the extreme corners of the plane and denote them \( I_f^+ I_s^+ \), \( I_f^- I_s^- \), \( I_f^- I_s^+ \), \( I_f^+ I_s^- \).

\( I_t \) has a small range of variation (97% of the candidates have \( I_t < 0.1 \)), nevertheless it can differentiate, as a third dimension, the expressions in the 4 groups defined above from a minimum to a maximum of interruptibility.

As one could presume, the expressions appearing in the group \( I_f^+ I_s^- \) with the lowest score of \( I_t \) are examples of opaque, crystalized or terminological expressions, such as testamento biologico ‘living will’ (\( I_f = 0.066, I_s = 0.004, I_t = 0 \)), valor militare ‘military valour’ (\( I_f = 0, I_s = 0, I_t = 0 \)), anidride carbonica ‘carbon dioxide’ (\( I_f = 0, I_s = 0, I_t = 0.001 \)). However expressions in the same group with the highest values of interruptibility\(^{12}\) seem to be compositional and just lexically restricted: carriera solista ‘solo career’ (\( I_f = 0.067, I_s = 0.018, I_t = 0.280 \)), sito ufficiale ‘official website’ (\( I_f = 0.043, I_s = 0.077, I_t = 0.076 \)).

Similar results come out for the group \( I_f^+ I_s^+ \), where expressions like cartone animato ‘cartoon’ (\( I_f = 0.333, I_s = 0.033, I_t = 0.0004 \)), macchina fotografica ‘camera’ (\( I_f = 0.374, I_s = 0.058, I_t = 0.004 \)), appear with low scores of interruptibility, while punto debole ‘weak point’ (\( I_f = 0.4, I_s = 0.066, I_t = 0.052 \)), figlio maschio ‘male son’ (\( I_f = 0.479, I_s = 0.098, I_t = 0.037 \)), have the highest values of interruptibility.

\(^{12}\)Recall that here, due to the high frequency of the expressions and to \( I_t \)’s range of variation, values of \( I_t \) close to 0.1 represent expressions that are sufficiently interrupted.

For \( I_f^- I_s^+ \), we have free combinations for higher \( I_t \), such as colore bianco ‘white colour’ (\( I_f = 0.097, I_s = 0.385, I_t = 0.129 \)) or colore rosso ‘red colour’ (\( I_f = 0.066, I_s = 0.362, I_t = 0.097 \)), and more lexically restricted expressions for lower values, such as corpo umano ‘human body’ (\( I_f = 0.077, I_s = 0.534, I_t = 0.008 \)), fama internazionale ‘international fame’ (\( I_f = 0.011, I_s = 0.441, I_t = 0.007 \)).

Finally the group \( I_f^+ I_s^- \) presents only expressions with very low values of \( I_t \) depending on the fact that expressions with high interruptibility, high substitutability and free inflection have been presumably excluded from the list because of their low AM scores. The remaining expressions in the group are of the kind of spettacolo teatrale ‘theatre performance’ (\( I_f = 0.468, I_s = 0.365, I_t = 0.006 \)), partito politico ‘political party’ (\( I_f = 0.471, I_s = 0.562, I_t = 0.003 \)), thus mainly compositional.

5 Discussion and Interpretation

By analysing the distribution of MWE candidates, it is possible to consider the scheme of Table 1 in which the following three categories appear: free combinations, multiword units and lexical collocations. As one can note, inflection variability does not play a role in discriminating between the categories.

It must be underlined that the three indices group the expressions into sets that appear to be more or less homogeneous with respect to the intuitive distinction between semantic units and compositional, lexically restricted expressions.

Free combinations represent the “false positives” of the list, i.e. expressions that do not need a special categorical status in phraseology.

Multiword units (polirematiche) represent here a subcategory of MWEs which exhibit the fol-
loving features: they can be metaphoric ("catena montuosa ‘mountain chain’"), completely crystallized ("quartier generale ‘headquarter’"), terminological ("amministratore delegato ‘managing director’"), they can present an unpredictable semantic addition ("gas naturale, ‘natural gas’, meaning the gas provided in houses for domestic uses"), or one of the components assumes a specific and unusual meaning ("casa automobilistica ‘car company’, lit. ‘car house’"). Despite their variability, the entities in this group are all perceived as “units” of meaning because the lack of one of the components makes the expressions lose their overall meaning.

Finally, lexical collocations represent here those entities that are generally perceived as fully compositional, being “not fixed but recognizable phraseological units” (Tiberii, 2012). They exhibit the following possible features: one of the component is used only in combination with the other one ("acqua potabile ‘drinking water’, where potabile only refers to water"), or although other synonymous words are available and could give the expression the same meaning, just one specific component word is preferred ("sito ufficiale ‘official site’"; cf. "sito autorizzato").

### 6 Further considerations and limits

Although not reported here, expressions with values for $I_f, I_s \in [0.1, 0.33]$ show continuity between the categories of Table 1. Moreover, since our thesaurus does not deal with sense disambiguation, a manual check on concordances was performed. For very few metaphorical expressions, $I_s$ produced non-reliable values, since it can happen that, once a synonym of one component has been substituted for the original word, the new expression is still highly attested in the corpus, although it has lost the original metaphorical meaning. In order to correct this bias in the future, the criterion of substitutability should check, for example, not only the number of attested replaced expressions, but also if they share the same context words of the basic expression.

### 7 Conclusion and future work

Our analysis shows that the intuitive distinction between the two main subcategories of MWEs (multiword units vs. lexical collocations) can be empirically reproduced by testing the syntactic and semantic behaviour of the expressions on corpora. In this way we provide an empirical criterion, related to the intuitive and hardly definable notion of compositionality, able to attest how expressions exhibit different restrictions depending on their subcategory. Multiword units are characterized by low values of interruptibility and low values of substitutability. Lexical collocations can be more easily interrupted if they have low values of substitutability, while they do not allow for interruptibility if they have high substitutability. Since also a subgroup of free combinations is identified when intersecting the values of the indices, our methodology can be used as well for automatic removal of false positives from MWE candidate lists.

Future work must include the extension of the analysis to other forms of nominal MWEs as well as other grammatical categories by the development of tools which can deal with verbal or adverbial MWEs, as well as tests on different corpora.

---

13This consideration relates our work to that of Baldwin et al. (2003), Bannard (2007), Weller and Fritzinger (2010), Cap et al. (2013), whose goal is to implement the identification of true positive candidates by using both syntactic or semantic features and AMs.
References
Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment (ACL 2003), pages 89–96.

Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In Proceedings of the Workshop on a Broader Perspective on Multiword Expressions, pages 1–8.

Sabine Bartsch. 2004. Structural and Functional Properties of Collocations in English. Narr, Tübingen.

Ignacio Bosque. 2004. Combinatoria y significación. Algunas reflexiones. In REDES, Diccionario Combinatorio del Español Contemporaneo. Hoepli.

Fabienne Cap, Marion Weller, and Ulrich Heid. 2013. Using a Rich Feature Set for the Identification of German MWEs. In Proceedings of Machine Translation Summit XIV, Nice, France.

Federica Casadei. 1996. Metafore ed Espressioni Idiomatichie. Uno studio semantico sull’italiano. Bulzoni Editore, Roma.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1):22–29.

Tullio De Mauro. 2007. GRADIT, Grande Dizionario Italiano dell’Uso. UTET.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1):61–74.

Stefan Evert. 2004. The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph.D. thesis, University of Stuttgart.

Stefan Evert. 2008. Corpora and Collocations. In A. Lüdeling and M. Kytö, editors, Corpus Linguistics. An International Handbook. Mouton de Gruyter, Berlin.

Afsaneh Fazly and Suzanne Stevenson. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In Proceedings of the Workshop on A Broader Perspective on Multiword Expressions. ACL, pages 9–16.

J. R. Firth. 1957. Papers in Linguistics. Oxford University Press, Oxford.

Patrick Hanks. 2013. Lexical Analysis. MIT Press, Cambridge, MA.

Francesca Masini. 2007. Parole sintagmatiche in italiano. Ph.D. thesis, Università degli Studi di Roma Tre.

Malvina Nissim and Andrea Zaninello. 2011. A quantitative study on the morphology of Italian multiword expressions. Lingue e linguaggio, (2):283–300.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a Framework for Multiword Expression Identification. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta.

Paola Tiberii. 2012. Dizionario delle collocazioni. Le combinazioni delle parole in italiano. Zanichelli.

Marion Weller and Fabienne Fritzinger. 2010. A hybrid approach for the identification of multiword expressions. In Proceedings of the SLTC 2010 Workshop on Compounds and Multiword Expressions.