Extracting MWEs from Italian corpora:  
A case study for refining the POS-pattern methodology  

Malvina Nissim  
FICLIT  
University of Bologna  
malvina.nissim@unibo.it  

Sara Castagnoli  
LILEC  
University of Bologna  
s.castagnoli@unibo.it  

Francesca Masini  
LILEC  
University of Bologna  
francesca.masini@unibo.it

Abstract
An established method for MWE extraction is the combined use of previously identified POS-patterns and association measures. However, the selection of such POS-patterns is rarely debated. Focusing on Italian MWEs containing at least one adjective, we set out to explore how candidate POS-patterns listed in relevant literature and lexicographic sources compare with POS sequences exhibited by statistically significant n-grams including an adjective position extracted from a large corpus of Italian. All literature-derived patterns are found—and new meaningful candidate patterns emerge—among the top-ranking trigrams for three association measures. We conclude that a final solid set to be used for MWE extraction will have to be further refined through a combination of association measures as well as manual inspection.

1 Introduction
The CombiNet project has the goal of building an online resource for Word Combinations in Italian, including MWEs of various degrees of fixedness (such as phrasal lexemes, collocations and usual combinations) as well as distributional profiles of Italian lexemes. Within this project, the present paper aims at investigating ways to refine a well-known methodology for MWE-extraction, namely the combined use of previously identified POS-patterns and association measures (Evert and Krenn, 2005). While POS-patterns are widely used to extract MWEs from corpora in order to constrain the array of possible outputs (Krenn and Evert, 2001; Wermter and Hahn, 2006, e.g.), the way in which POS-patterns are created in the first place is much less addressed. This step is however crucial, especially considering that the list of patterns is necessarily language-specific. The goal of this paper is to propose a method to optimize—in terms of both recall and precision—the list of POS patterns to be used for the subsequent extraction of potential MWEs. In order to do this, we compare predetermined patterns, which would be normally used as a first-pass sieve for potential MWEs, with patterns exhibited by statistically significant n-grams extracted from data.

2 Methodology
In this pilot study, we focus on MWEs containing at least one adjective, and we limit the extraction to trigrams (Section 2.1). We make use of the following sets of data: (a) a list of frequently used Italian adjectives; (b) a list of previously identified POS-patterns containing at least one adjective.

The adjectival lemmas were taken from the Senso Comune dictionary, which contains 2,010 fundamental lemmas of the Italian lexicon, 211 of which are adjectives (e.g. *bello* “beautiful”, *brutto* “ugly”, *ricco* “rich”). These adjectives are used to constrain the extraction procedure, and we refer to this set as \{SC\}.

The list of predetermined POS-patterns for MWEs involving one adjective was obtained by merging the following information: (a) patterns of word combinations included in existing combinatory dictionaries for Italian (Piunno et al., 2013), see Table 1a; (b) additional combinatory types mentioned in the relevant theoretical literature (Voghera, 2004; Masini, 2012), summarised in Table 1b; and (c) a few more patterns based on our own intuition, i.e. identified by elaborating on the previous two lists (Table 2). This joint collection contains a total of 19 patterns, most of which are bigrams (11), and fewer are trigrams (8). Note

---

1https://sites.google.com/site/enwcin/home
2For information on POS tags see Appendix.
3http://www.sensocomune.it/
that trigrams (put together in Table 2) come for the
most part from our intuition, indicating that these
patterns are rather neglected in the literature and
in combinatory dictionaries of Italian, which tend
to focus on bigrams. For this reason, and because
longer sequences are intrinsically more idiosyn-
cratic, we concentrate on trigrams for this pilot
experiment, although in the discussion we take into
account bigrams, too (Section 3).

Table 1: Italian POS-patterns with ADJ(s)

| POS-pattern       | Example                             | Translation                      |
|-------------------|-------------------------------------|----------------------------------|
| (a) from lexicographic sources |
| ADJ ADJ           | stanco morto                        | dead tired                       |
| ADJ CON ADJ       | vivo e vegeto                       | live and kicking                 |
| ADJ NOUN          | prima classe                        | first class                      |
| ADJ PRE           | pronto a                            | ready to                         |
| ADV ADJ           | molto malato                        | very ill                         |
| NOUN ADJ          | casa editrice                       | publishing house                 |
| VER ADJ           | uscire pazzo                        | to go crazy                       |
| (b) from relevant literature |
| ADJ PRO           | qual esso                           | which/who                        |
| ADJ, ADJ, ADJ     | papale papale                       | bluntly                          |
| ART PRE ADJ       | alla francese                       | French-style                     |
| PRE ADJ           | a caldo                             | on the spot                      |
| PRE ADJ NOUN      | di bassa lega                       | vulgar/coarse                    |
| PRE NOUN ADJ      | a senso unico                       | one-way                          |
| PRO ADJ           | tal altro                           | some other                       |

2.1 Extracting the trigrams

From the corpus La Repubblica (Baroni et al.,
2004), which consists of 300M words of newswire
contemporary Italian, we extracted all trigrams fea-
turing at least one adjective, deriving this informa-
tion from the pre-existing POS tags in the corpus.
All trigrams were extracted as sequences of lems-
as. We created three separate lists according to
the adjective’s position in the trigram (first, second,
or third). All instances containing any punctuation
item were discarded.

For each of the three sets, we kept only trigrams
occurring more than five times in the whole cor-
pus. As a further step, we selected those instances
featuring one of the 211 adjectives in \{SC\}, yield-
ing a total of 89,217 different trigrams featuring
an adjective as first member (191 adjectives from
\{SC\} occur in this position), 100,861 as second
(192 adjectives), and 114,672 as third (193).

2.2 Ranking the trigrams

We used the Text-NSP package (Banerjee and Ped-
ersen, 2003) to rank the trigrams in each of the
three sets according to three association measures
(AMs), namely the Poisson-Stirling measure (PS),
the log-likelihood ratio (LL) and pointwise mutual
information (PMI). However, on the basis of
preliminary inspection and observations in the liter-
ature on ranking Italian MWEs extracted from cor-
pora (Nissim and Zaninello, 2013), we discarded
PMI as not too accurate for this task. We also
considered raw frequencies, as they have proved
good indicators for collocations, on a par with AMs
(Krenn and Evert, 2001; Bannard, 2007).

The idea is to check which POS sequences are
exhibited by the highest instances in the rank, un-
der the rationale that such patterns might be good
representations of Italian MWEs containing adjectives,
and can be used for further extraction and characterisation of the phenomenon (in dictionar-
ies and resources). Thus, we selected the top 10% instances in each rank, extracted their POS patterns,
and ranked such patterns according to the number
of times they appeared. Tables 3–5 report the ten
most frequent patterns according to each measure,
when an adjective is found in first, second, and
third position, respectively.

3 Analysis and discussion

By comparing the ranked patterns in Tables 3–5
with the predetermined POS-patterns for trigrams
in Table 2, we draw the following observations.

We first consider patterns that are ranked high
for all measures. Some find a correspondence to
those in Table 2, implying that these are likely to
be solid, characteristic POS sequences to be used
in extraction (ADJ CONJ ADJ (for ADJ in first
position), ADJ PRE VER, PRE ADJ NOUN, and
VER PRE ADJ). Other found patterns, instead, are
not among the pre-identified ones, but are definitely
typical sequences, as the analysis of some of the
extracted trigrams shows. Among these: ADJ PRE
NOUN (ospite d’onore “special guest”), VER ART
ADJ (essere il solo “to be the only one”), NOUN
PRE ADJ (agente in borghese “plain-clothes po-
lieutenant”), ART PRE ADJ NOUN (all’ultimo mo-
mento “at the last moment”). Envisaging an ex-
traction procedure based on POS sequences, such
structures should be included to improve recall.

Conversely, the PRE ART ADJ pattern exhibits
an incomplete sequence, and is therefore unsuitable
for MWE extraction. Since the inclusion of such
patterns would possibly affect precision, they need
to be filtered out on the grounds of grammatical
Table 2: Trigram POS-patterns containing ADJ(s)

| POS-pattern       | Example Translation from literature and resources |
|-------------------|---------------------------------------------------|
| ADJ CON ADJ       | pura e semplice                                   |
| PRE ADJ NOUN      | a breve termine                                   |
| PRE NOUN ADJ      | in tempo reale                                    |

|                          | from our own intuition                           |
|-------------------------|-------------------------------------------------|
| ADJ PRE VER             | duro a morire                                    |
| NOUN NOUN ADJ           | prodotto interno lordo                           |
| NOUN NOUN ADJ           | dipartimento affari sociali                      |
| PRE ADJ VER             | per quieto vivere                                |
| VER PRE ADJ             | dare per scontato                                |

Table 3: Top 10 POS patterns featuring an adjective as word1, extracted from the top 10% trigrams ranked according to LL, PS, and raw frequency.

| LL                      | PS                      | raw frequency |
|-------------------------|-------------------------|---------------|
| ADJ PRE VER             | ADJ NOUN PRE            | ADJ NOUN PRE  |
| ADJ PRE ART             | ADJ NOUN ARTPRE         | ADJ NOUN ARTPRE |
| ADJ NOUN PRE            | ADJ NOUN ADJ            | ADJ NOUN ADJ  |
| ADJ PRE NOUN            | ADJ ARTPRE NOUN         | ADJ ARTPRE NOUN |
| ADJ NOUN ARTPRE         | ADJ PRE VER             | ADJ PRE VER  |
| ADJ ARTPRE NOUN         | ADJ PRE NOUN            | ADJ PRE NOUN  |
| ADJ PRE DET             | ADJ CON ADJ             | ADJ CON ADJ  |
| ADJ CON ADJ             | ADJ NPR NPR             | ADJ CON ADJ  |
| ADJ CHE CLI             | ADJ NOUN CON            | ADJ NOUN CON  |
| ADJ DET NOUN            | ADJ PRE ART             | ADJ CON ART  |

Additionally, there are patterns that contain or are portions of more relevant patterns for MWE-ness. Some capture what are in fact bigrams (Table 6), while others are portions of 4-grams or possibly larger sequences, namely NOUN ADJ PRE (NOUN), (NOUN) ADJ ARTPRE NOUN, and NOUN ARTPRE ADJ (NOUN), where the “missing” POS is given in brackets. Examples are: concorso esterno in (omicidio) “external participation in (murder)”, (banca) nazionale del lavoro “National (Bank) of Labour”, and paese del terzo (mondo) “third world (country)”, respectively. Running a full-scale extraction procedure that accounts for all n-grams should naturally take care of this.

Some of patterns from Table 2 are ranked high only by some measures: PRE NOUN ADJ only according to PS and raw frequency (Table 5), and NOUN ADJ ADJ both for second and third position, but only by PS. Overall, with respect to their ability to extract previously identified POS-patterns, AMs perform similarly when the adjective is the first member (Table 3), whereas PS seems to be more indicative when the adjective is second and third (Tables 4-5), together with raw frequency, while LL seems to be generally performing the worst. This point calls for a combination of AMs (Pecina, 2008), but will require further work.

As for predetermined patterns that are not found among the top ones, we observe that NOUN NOUN ADJ is basically an adjective modifying a noun-noun compound, and should be best treated as a “complex bigram”. Similarly, the PRE ADJ VER pattern can be seen as an extension of the ADJ VER bigram, which is usually not considered (Table 1). Investigating the combination of bigrams, trigrams and n-grams with n > 3 is left for future work.

4 Conclusion

In summary, basically all of the literature/intuition-based patterns are retrieved from highly ranked plain trigrams. However, top-ranking trigrams also exhibit other POS sequences which should be included in a set of patterns used for MWE extrac-
tion to improve recall. At the same time, several patterns extracted with this technique are to be discarded. Some are just irrelevant (e.g. ADJ CHE CLI, nero che le “black that them”): in this respect, combining various AMs or setting grammatical constraints could help refine precision, but human intervention also seems unavoidable. Others are not meaningful trigrams as such, but may be meaningful as parts of larger MWEs or because they contain meaningful bigrams. Here, it would be interesting to explore how to combine n-grams with different n-values.

This pilot experiment shows that trigram ranking is useful to extract new patterns that are not considered in the initial set. The latter can be therefore expanded by following the proposed methodology, as a preliminary step towards the actual extraction of candidate MWEs from corpora. Clearly, the validity of the expanded POS-pattern set can only be evaluated after the extraction step is completed.
Acknowledgments

This research was carried out within the PRIN project 2010-2011 (n. 20105B3HE8_003) Word Combinations in Italian: theoretical and descriptive analysis, computational models, lexicographic layout and creation of a dictionary, coordinated by Raffaele Simone (Roma Tre University) and funded by the Italian Ministry of Education, University and Research (MIUR). We are grateful to the anonymous reviewers for their useful comments.

References

Satanjee Banerjee and Ted Pedersen. 2003. The design, implementation, and use of the ngram statistics package. In A. F. Gelbukh, editor, volume 2588 of Lecture Notes in Computer Science, pages 370–381.

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

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. In Proc. of LREC 2004, pages 1771–1774.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. Computer Speech & Language, 19(4):450–466. Special issue on Multiword Expression.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In Proc. of the ACL-EACL Workshop on Collocations, pages 39–46, Toulouse.

Francesca Masini. 2012. Parole sintagmatiche in italiano. Caissa, Roma.

Malvina Nissim and Andrea Zaninello. 2013. Modeling the internal variability of multiword expressions through a pattern-based method. ACM Trans. Speech Lang. Process., 10(2):7:1–7:26.

Pavel Pecina. 2008. A Machine Learning Approach to Multiword Expression Extraction. In Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions, pages 54–57, Marrakech, Morocco. European Language Resources Association.

Valentina Piunno, Francesca Masini and Sara Castagnoli. 2013. Studio comparativo dei dizionari combinatori dell’italiano e di altre lingue europee. CombiNet Technical Report. Roma Tre University and University of Bologna.

Miriam Voghera. 2004. Polirematiche. In Grossmann, Maria & Franz Rainer, editors, La formazione delle parole in italiano, Tübingen, Max Niemeyer Verlag, 56-69.

Joachim Wermter and Udo Hahn. 2006. You can’t beat frequency (unless you use linguistic knowledge): A qualitative evaluation of association measures for collocation and term extraction. In Proc. of COLING-ACL ’06, pages 785–792, USA.

Appendix

The tagset for all patterns extracted from the corpus “La Repubblica” is accessible at http://sslmit.unibo.it/~baroni/collocazioni/itwac.tagset.txt. In the context of this experiment we collapsed all fine-grained tags into the corresponding coarse-grained tag (e.g. all verbal tags such as VER:fin or VER:ppast were collapsed into VER). The POS tags used in this paper are to be interpreted as in the Table below.

| abbreviation | part of speech                                      |
|--------------|-----------------------------------------------------|
| ADJ          | adjective                                           |
| ADV          | adverb                                              |
| ART          | article                                             |
| ARTPRE       | prepositional article                                |
| CHE          | any function of the word “che” (adjective, conjunction, pronoun) |
| CLI          | clitic                                              |
| DET          | determiner                                          |
| NOUN         | noun                                                |
| PRE          | preposition                                         |
| VER          | verb                                                |

Table 7: POS tags used in this paper.