A Repository of Variation Patterns for Multiword Expressions

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

Andrea Zaninello
Zanichelli Editore, Humanities Department
andrea.zaninello@gmail.com

1 Introduction and Background

One of the crucial issues in the analysis and processing of MWEs is their internal variability. Indeed, the feature that mostly characterises MWEs is their fixedness at some level of linguistic analysis, be it morphology, syntax, or semantics. The morphological aspect is not trivial in languages which exhibit a rich morphology, such as Romance languages.

The issue is relevant in at least three aspects of MWE representation and processing: lexicons, identification, and extraction (Calzolari et al., 2002). At the lexicon level, MWEs are usual stored as one form only, the so-called *quotation form* (or *citation form*). However, *some* variations of the quotation form might also be valid instances of MWEs (Bond et al., 2005) — some but not all, as some of them might actually be plain compositional phrases.

This becomes relevant for automatic identification and extraction. If a lexicon stores the quotation form only, identification on a corpus done via matching lexicon strings as such would miss valid variations of a given MWE. Identification could be done exploiting lemmas rather than quotation forms, but an unrestricted match would also possibly return compositional phrases. Extraction is usually done applying association measures over instances of given POS patterns (Evert and Krenn, 2005), and because lemmas are matched, no restrictions on internal variation is enforced as such. Knowing *which* variations should be allowed for the quotation form of a given MWE would help in increasing recall while keeping precision high. However, specifying such variations for *each* MWE would be too costly and wouldn't help in extraction, as no specifications could be done a priori on yet unknown MWEs. Optimally, one would need to find more general variation patterns that could be applied to *classes* of MWEs. Indeed, the main idea behind this work is that MWEs can be handled through more general patterns. This is also claimed, for instance, by Masini (2007) whose analysis on Italian MWEs takes a constructionist perspective (Goldberg, 2003), by Weller and Heid (2010), who treat verbal expressions in German, and also by Grégoire (2010), who bases his work on the Equivalence Class Method (ECM, (Odijk, 2004)) assuming that MWEs may be clustered according to their *syntactic* pattern and treated homogeneously. We suggest that variation patterns can be found and defined over POS sequences. Working on Italian, in this paper we report the results of ongoing research and show how such patterns can be derived, we then propose a way to encode them in a repository, which can be combined with existing lexicons of MWEs. For the moment, we restrict our study to contiguous MWEs although we are aware that non-contiguous expressions are common and should be treated, too (see also (Pianta and Bentivogli, 2004)). Thus, only morphological variation is considered at this stage, while phenomena such as insertion and word order variation are left for future work.

2 Obtaining Variation Patterns

Variation patterns refer to POS sequences and rely on frequencies. The main resources needed for obtaining them are a MWE lexicon and a reference corpus (pos-tagged and lemmatised).We use a MWE lexicon derived from an existing online dictionary

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 101–105,
Atlanta, Georgia, 13-14 June 2013.

for Italian (Zaninello and Nissim, 2010), and the corpus "La Repubblica" (Baroni et al., 2004) for obtaining frequencies.

A *variation pattern* encodes the way a given instance of a MWE morphologically differs from its original quotation form in each of its parts. All tokens that correspond to the quotation form are marked as fix whereas all tokens that do not are marked as $flex$. Consider Example (1):

(1) a. quotation form: "casa di cura" (nursing home)
    b. instance: "case di cura" (nursing homes)
    c. variation pattern: $flex\_fix\_fix$

The pattern for the instance in (1b) is $flex\_fix\_fix$ because the first token, "case" (houses) is a plural whereas the quotation form features a singular ("casa", house), thus is assigned a $flex$ label, whereas the other two tokens are found exactly as they appear in the quotation form, and are therefore labelled as fix.

At this point, it is quite important to note that a binary feature applied to each token makes *flexibility* underspecified in at least two ways. First, the value $flex$ does not account by itself for the degree of variation: a token is $flex$ if it can be found in one variation as well as many. We have addressed this issue elsewhere via a dedicated measure (Nissim and Zaninello, 2011), but we do not pick it up here again. In any case, the degree of variation could indeed be included as additional information. Second, we only specify which part of the MWEs varies but do not make assumptions on the type of variation encountered (for example, it doesn't distinguish at the level of gender or number).

We believe this is a fair tradeoff which captures generalisations at a level which is intermediate between a word-by-word analysis and considering the entire MWE as a single unit. Additionally, it does not require finer-grained annotation than POS-tagging and lemmatisation, and allows for the discovery of possibly unknown and unpredicted variations. Morphological analysis, when needed, is of course still possible *a posteriori* on the instances found, but it is useful that at this stage flexibility is left underspecified.

As said, validating variation patterns per MWE would be impractical and uninformative with respect to the extraction of previously unseen MWEs. Thus, we define variation patterns over part-of-speech sequences. More specifically, we operate as follows:

1. search all MWEs contained in a given lexicon on a large corpus, matching all possible variations (lemma-based, or unconstrained, search);
2. obtain variation patterns for all MWEs by comparing each instance to its quotation form;
3. group all MWEs with the same POS sequence;
4. for each POS sequence collect all variation patterns of all pertinent MWEs.

In previous work (Nissim and Zaninello, 2013), we have observed that frequency is a good indicator of valid patterns: the most frequent variation patterns correlate with variations annotated as correct by manual judges. Patterns for two nominal POS were evaluated, and they were found to be successful. In this paper we pick three further POS sequences per expression type for a total of nine POS patterns, and evaluate the precision of a pattern selection measure.

The availability of variation patterns per POS sequences (and expression type) can be of use both in identification as well as in extraction. In identification, patterns can be used as a selection strategy for all of the matched instances. One could just use frequency directly from the corpus where the identification is done, but this might not always be possible due to corpus size. This is why using an external repository of patterns evaluated against a large reference corpus for a given language might be useful.

In extraction tasks, patterns can be used as filters, either as a post-processing phase after matching lemmas for given POS sequences, or directly extracting only allowed configurations which could be specified for instance in extraction tools such as `mwetoolkit` (Ramisch et al., 2010). In previous work we have shown that patterns can be derived comparing found instances against their lemmatised form, making this a realistic setting even in extraction where quotation forms are not known (Nissim and Zaninello, 2013).

3 Ranking

For ranking variation patterns we take into account the following figures:

- the total number of different variation patterns per POS sequence
- the total number of instances (hits on the corpus) with a given variation pattern

For example, the POS sequence ADJ_PRE_NOUN characterising some adjectival expressions is featured by 9 different original multiword expressions that were found in the corpus. The variations with respect to the quotation form (indicated as fix_fix_fix and found for seven different types) in which instances have been found are four: $flex\_fix\_fix$ (13 times), $flex\_fix\_flex$ (7 times), fix_fix_flex (3 times), and fix_flex_flex (one time), for a total of 31 variations. Each instance yielding a given pattern was found at least once in the corpus, but possibly more times. We take into account this value as well, thus counting the number of single *instances* of a given pattern. So, while "degni di nota" ("worth$_{pl}$ mentioning", quotation form: "degno di nota", "worth$_{sg}$ mentioning") would serve as *one* variation of type $flex\_fix\_fix$, counting instances would account for the fact that this expression was found in the corpus 38 times. For the ADJ_PRE_NOUN sequence, instances of pattern fix_fix_fix were found 130 times, instances of $flex\_fix\_fix$ 219, $flex\_fix\_flex$ 326, fix_fix_flex 90, and fix_flex_flex just once, for a total of 766 instances.

Such figures are the basis for pattern ranking and are used in the repository to contribute to the description of variation patterns (Figure 1). We use the share of a given variation pattern (vp) over the total number of variations (*pattern share*). In the example above, the share of $flex\_fix\_fix$ (occurring 13 times) would be 13/31 (41.9%), as 31 is the total of encountered variations for the ADJ_PRE_NOUN POS sequence. We also use the *instance share*, which for the same variation pattern would be 219/766 (12.0%) and combine it with the pattern share to obtain an overall share ($share_{vp}$):

$$share_{vp} = (\frac{\#variations_{vp}}{\#variations_{pos}} + \frac{\#instances_{vp}}{\#instances_{pos}})/2$$

As a global ranking score (GRS_{vp}), the resulting average share is combined with the *spread*, namely the ratio of instances over variations (219/13 for $flex\_fix\_fix$), a pattern-internal measure indicating the average instances per variation pattern.

$$spread_{vp} = \frac{\#instances_{vp}}{\#variations_{vp}}$$

$$GRS_{vp} = share_{vp} * spread_{vp}$$

Only patterns with $GRS > 1$ are kept, with the aim of maximising precision. Evaluation is done against some POS sequences for which extracted instances have been manually annotated. Precision, recall, and f-score are reported in Table 1. Results for an unconstrained search (no pattern selection) are also included for comparison. The number of variation patterns that we keep on the basis of the ranking score includes the fix_fix_fix pattern.

From the table, we can see that in most cases precision is increased over an unconstrained match. However, while for verbal expressions the boost in precision preserves recall high, thus yielding f-scores that are always higher than for an unconstrained search, the same isn't true for adjectives and adverbs. In two cases, both featuring the same POS sequence (PRE_NOUN_ADJ) though for different expression types, recall is heavily sacrificed. In three cases, the GRS doesn't let discard any patterns, thus being of no use in boosting precision. These are cases where only two variation patterns were observed, indicating that possibly other ranking measures could be explored for better results under such conditions. In previous work we have seen that selecting variation patterns works well for nominal expressions (Nissim and Zaninello, 2013).

Overall, even though in some cases our method does not yield different results than an unconstrained search, whenever it does, precision is always higher. It is therefore worth applying whenever boosting precision is desirable.

4 Repository and Encoding

We create an XML-based repository of POS patterns with their respective variation patterns. Variation patterns per POS sequence are reported according to the ranking produced by the GRS. However, we

Table 1: Evaluation of pattern selection for some POS sequences according to the Global Ranking Score.

		GRS				unconstrained		
expr type	POS sequence	# vp kept	prec	rec	f-score	prec	rec	f-score
verbal	VER:infi_ARTPRE_NOUN	2/4	1.000	0.998	0.999	0.979	1.000	0.989
	VER:infi:cli_ART_NOUN	2/7	0.965	0.981	0.973	0.943	1.000	0.971
	VER:infi_ADV	2/4	0.997	0.978	0.987	0.951	1.000	0.975
adjectival	ADJ_PRE_NOUN	2/2	0.379	1.000	0.550	0.379	1.000	0.550
	PRE_NOUN_ADJ	1/4	1.000	0.590	0.742	0.848	1.000	0.918
	PRE_VER:fin	4/5	1.000	0.968	0.984	1.000	1.000	1.000
adverbial	PRE_ADV	2/2	0.671	1.000	0.803	0.671	1.000	0.803
	PRE_NOUN_ADJ	1/4	1.000	0.746	0.854	0.899	1.000	0.947
	PRE_ADJ	2/2	0.362	1.000	0.532	0.362	1.000	0.532

include all observed patterns equipped with the frequency information we used, so that other ranking measures or different thresholds could be applied.

The repository is intended as connected to two sources, namely a lexicon to obtain quotation forms of MWEs to be searched, and the corpus where expressions were searched, which provides the figures.

POS patterns are listed as elements for each `expression` element, whose attribute `type` specifies the grammatical type—for example "verbal". The same POS pattern can feature under different expression types, and could have different constraints on variation according to the grammatical category of the MWE (in extraction this issue would require dedicated handling, as the grammatical category is not necessarily known in advance). For the element `pattern`, which specifies the POS sequence, the attribute `mwes` indicates how many different original mews were found for that sequence, and the attributes `variations` and `instances` the number of variations and instances (Section 3). Actual patterns are listed as data of a `vp` (variation pattern) element, according to decreasing GRS, with values obtained from the reference corpus (specified via a `corpus` element). Attributes for the `vp` element are `vshare` (variation share), `ishare` (instance share), `spread`, and `grs` (see again Section 3). In Figure 1 we provide a snapshot of what the repository looks like.

The POS sequence of a MWE in the original lexicon can be matched to the same value in the repository, and so can the expression type, which should also be specified in the lexicon, so that the relative variation patterns can be inherited by the MWE.

References

M. Baroni, S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston, and M. Mazzoleni. 2004. Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. In *Proceedings of LREC 2004*, pages 1771–1774.

F. Bond, A. Korhonen, D. McCarthy, and A. Villavicencio. 2005. Multiword Expressions: Having a crack at a hard nut. *Computer Speech and Language*, 19:365–367.

N. Calzolari, C. J. Fillmore, R. Grishman, N. Ide, A. Lenci, C. MacLeod, and A. Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of LREC 2002*, pages 1934–1940.

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 Expressions.

Adele Goldberg. 2003. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224.

Nicole Grégoire. 2010. DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44(1-2):23–39.

Francesca Masini. 2007. *Parole sintagmatiche in italiano*. Ph.D. thesis, University of Roma Tre, Rome, Italy.

Malvina Nissim and Andrea Zaninello. 2011. A quantitative study on the morphology of Italian multiword expressions. *Lingue e Linguaggio*, X:283–300.

Malvina Nissim and Andrea Zaninello. 2013. Modelling the internal variability of multiword expressions through a pattern-based method. *ACM Transactions on Speech and Language Processing*, Special issue on Multiword Expressions.

```
<corpus name="larepubblica">
 <expression type="verbal">
  <patterns>
   <pattern pos="VER:infi_ARTPRE_NOUN" mwes="55" variations="671" instances="9046">
    <vp vshare="0.896" ishare"0.740" spread="42.1" grs="9.109">flex_fix_fix</vp>
    <vp vshare="0.082" ishare"0.256" spread="11.1" grs="7.127">fix_fix_fix</vp>
    <vp vshare="0.016" ishare"0.003" spread="2.6" grs="0.026">flex_flex_fix</vp>
    <vp vshare="0.006" ishare"0.000" spread="1.2" grs="0.004">flex_flex_flex</vp>
   </pattern>
   <pattern pos="VER:infi:cli_ART_NOUN" mwes="41" variations="600" instances="3703">
    <vp vshare="0.065" ishare"0.267" spread="25.3" grs="4.203">fix_fix_fix</vp>
    <vp vshare="0.893" ishare"0.723" spread="5" grs="4.040">flex_fix_fix</vp>
    <vp vshare="0.030" ishare"0.008" spread="1.6" grs="0.029">flex_flex_flex</vp>
    <vp vshare="0.005" ishare"0.000" spread="1" grs="0.003">flex_flex_fix</vp>
    <vp vshare="0.003" ishare"0.000" spread="1" grs="0.002">fix_flex_flex</vp>
    <vp vshare="0.002" ishare"0.000" spread="2" grs="0.002">fix_flex_fix</vp>
    <vp vshare="0.002" ishare"0.000" spread="1" grs="0.000">flex_fix_flex</vp>
   </pattern>
   <pattern ...>
    ...
   </pattern>
  </patterns>
 </expression>
 <expression type="adverbial">
  <patterns>
   <pattern pos="PRE_NOUN_ADJ" mwes="53" variations="79" instances="12202">
    <vp vshare="0.671" ishare"0.989" spread="227.7" grs="189.0">fix_fix_fix</vp>
    <vp vshare="0.076" ishare"0.007" spread="14" grs="0.580">fix_flex_flex</vp>
    <vp vshare="0.190" ishare"0.004" spread="2.9" grs="0.284">fix_fix_flex</vp>
    <vp vshare="0.063" ishare"0.000" spread="1" grs="0.032">fix_fix_fix</vp>
   </pattern>
   ...
  </patterns>
 </expression>
 <expression type="adjectival">
  <patterns>
   ...
  </patterns>
 </expression>
</corpus>
```

Figure 1: Snapshot of the XML repository of variation patterns over POS patterns, listed by expression types. See text for element and attribute explanation..

J. Odijk. 2004. A proposed standard for the lexical representation of idioms. In *Proceedings of EURALEX 2004*, pages 153–164.

Emanuele Pianta and Luisa Bentivogli. 2004. Annotating discontinuous structures in xml: the multiword case. In *Proceedings of LREC Workshop on XML-based Richly Annotated Corpora*, pages 30–37, Lisbon, Portugal.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a framework for multiword expression identification. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Marion Weller and Ulrich Heid. 2010. Extraction of German Multiword Expressions from Parsed Corpora Using Context Features. In *Proceedings of the seventh conference on International Language Resources and Evaluation (LREC 2010)*, pages 3195–3201. European Language Resources Association.

Andrea Zaninello and Malvina Nissim. 2010. Creation of Lexical Resources for a Characterisation of Multiword Expressions in Italian. In *Proceedings of LREC 2010*, pages 655–661, Valletta, Malta, may. European Language Resources Association (ELRA).