Creation of lexical resources for a characterisation of multiword expressions in Italian

Andrea Zaninello, Malvina Nissim

Dipartimento di Studi Linguistici e Orientali
Alma Mater Studiorum—Università di Bologna
andrea.zaninello@studio.unibo.it,malvina.nissim@unibo.it

Abstract

The theoretical characterisation of multiword expressions (MWEs) is tightly connected to their actual occurrences in data and to their representation in lexical resources. We present three lexical resources for Italian MWEs, namely an electronic lexicon, a series of example corpora and a database of MWEs represented around morphosyntactic patterns. These resources are matched against, and created from, a very large web-derived corpus for Italian that spans across registers and domains. We can thus test expressions coded by lexicographers in a dictionary, thereby discarding unattested expressions, revisiting lexicographers's choices on the basis of frequency information, and at the same time creating an example sub-corpus for each entry. We organise MWEs on the basis of the morphosyntactic information obtained from the data in an electronic, flexible knowledge-base containing structured annotation exploitable for multiple purposes. We also suggest further work directions towards characterising MWEs by analysing the data organised in our database through lexico-semantic information available in WordNet or MultiWordNet-like resources, also in the perspective of expanding their set through the extraction of other similar compact expressions.

1. Introduction

The label "multiword expression" (MWE) denotes a heterogeneous group of linguistic expressions composed of two or more words functioning as a single unit with respect to some levels of analysis (Calzolari et al., 2002). This includes (semi-)fixed expressions, idioms, compound nominals, verb particle and light verb constructions, institutionalized phrases etc. The essential role played by MWEs in Natural Language Processing (NLP) and linguistic analysis in general has been long recognised, as confirmed by the numerous specifically dedicated workshops and special issues of journals in recent years (CSL, 2005; JLRE, 2009). Although addressed from very different perspectives, it seems clear that in order to develop more efficient systems for the automatic identification and handling of MWEs in lexical resources we need a "deeper understanding of the structural and semantic properties of MWEs, such as morpho-syntactic patterns, semantic compositionality, semantic behaviour in different contexts, cross-lingual transformation of MWE properties etc." (Rayson et al., 2009).

As a consequence, the need for thorough descriptive studies on MWEs in different languages, along with reliable gold standards and guidelines for construction, annotation and evaluation of linguistic resources proves essential, as does the need to make these resources (such as MWE databases, metalinguistic descriptions and mark-ups, dedicated corpora, lexicons etc.) available for other researchers (Calzolari et al., 2002; Bond et al., 2005).

MWEs lie at a crossroad between automatically obtained data from corpora (through exploitation of association measures, for instance, or detection of idiosyncratic behaviour at the morpho-syntactic level), and choices operated by lexicographers, often on the basis of their intuition and common knowledge, as also shown by the collection of studies in (Fellbaum, 2007). Both views have limitations, but we believe these could be alleviated by integrating a corpus-driven and a lexicographic approach. Thus, we started out our investigation of MWEs in Italian by matching and evaluating lexicographer choices against occurrences in a very large corpus.

More specifically, the short-term aims of the study described in this paper are:

- acquiring information from very large corpora regarding expressions that lexicographers have classified as MWEs and included in the dictionary as such; this allows us to evaluate whether their choices have correspondence in real-world occurring data, which is a variable to be taken into serious account during the design of lexicographic resources.
- designing and developing a set of lexical resources, namely i) an electronic lexicon of Italian MWEs that takes advantage of the information available in manually compiled dictionaries; ii) a series of example corpora including large sets of MWE instances in their authentic context; iii) a flexible database of MWEs organised around morphosyntactic patterns obtained through corpus crawling;
- validating the information encoded in the database by manual post-processing and providing guidelines for the exploitation of the resource for future research, such as expanding the set of MWEs with compact and similar expressions through corpus-crawling.

From a methodological point of view, the development of this work should put us in a position to address two main issues regarding the treatment of MWEs, with specific focus on Italian: (a) what should be included in electronic resources of various kinds (dictionaries, ontologies, etc.) and how it should be represented, and (b) how to acquire relevant information for the development of systems for the automatic detection of MWEs as well as for human use.

Moreover, viewing these findings in the light of established theoretical models generally developed for other languages (especially English) might allow one to discover distinctive properties of MWEs in Italian also due to structural and typological differences (e.g. the distinctive fact that Italian is a pro-drop language, that it exhibits a richer morphology compared to other languages such as English, etc.). Further studies based on the characterised data provided by the resource may also suggest that subcategories of MWEs (or different dimensions of categorization in general) do not work for all languages in the same way, leading to an improvement and possibly enrichment of the existing theoretical frameworks.

The remainder of this paper is organised as follows. In Section 2. we discuss the issues related to the creation of lexical resources for MWEs, and their importance for NLP tasks. The experiments on Italian and the resources that we develop are described in Section 3., whereas observations on data are in Section 4.. In Section 5. we discuss directions for future research.

2. MWEs and Lexical Resources

The growing interest in MWEs over the past two decades reflects the increasing awareness in the NLP community that tasks involving syntactic or lexico-semantic processing such as machine translation, information retrieval, question-answering, and word sense disambiguation, require a robust handling of these expressions to improve their performances.

Even though some expressions 'prototypically' define the set of MWEs in a language (cf. idioms, verb particle constructions, light verbs constructions etc.), the group is however better conceived of as a continuum where elements have different degrees of inclusion between two poles of the lexicon and the syntax, with fuzzy boundaries at best (cf. Masini 2007): extreme examples are frozen, syntactically irregular fixed expressions such as 'by and large', at one end of the spectrum, and compositional, transparent, although statistically marked expressions such as 'traffic lights' on the other.

In fact, various degrees of MWE-hood seem to depend on the interaction of a plurality of factors (Moon, 1998), including syntactic and lexical features (syntactic irregularity, lexico-grammatical fixedness, non-substitutability, degree of lexicalisation, etc.), semantic and pragmatic considerations (opaqueness, non-compostitionality, recoverability of meaning from surface structures, proverbiality, figuration etc.) as well as statistical markedness. However, it is not clear whether these parameters hold cross-linguistically or must be idiosyncratically specified for each language.

The idiosyncratic nature of MWEs not only makes it difficult to tackle computational tasks such as extraction, identification, analysis and classification, but also makes it extremely complex to devise appropriate criteria for organising them in (possibly even cross-linguistic) electronic resources (such as electronic dictionaries, lexical databases, thesauri and ontologies etc.), which can also be used for further exploitation by automated systems. One of the main challenges of dealing with MWEs in the design of lexical resources is finding the optimal balance between their (inter- and intra-linguistic, syntactic and lexical) variability, and the need for maximal generalisation (Calzolari et al., 2002; Copestake et al., 2002; Villavicencio et al., 2004; Bond et al., 2005).

In the past few years, a lot of descriptive work has been done on MWEs, initially focusing on English (Nunberg et al., 1994; Moon, 1998; Sag et al., 2002, e.g.) although progressively more attention has been given to MWEs in other languages. For example, as (Rayson et al., 2009) points out, extensive work in this direction has been carried out at Lancaster University, resulting in a large collection of semantically annotated English, Finnish and Russian MWE lexical resources for a semantic annotation tool (Rayson et al., 2004; Piao et al., 2006). Efforts have been made to find efficient and portable ways to organise MWEs and MWE-related knowledge in lexical resources to be used in lexicography and NLP, also through corpus evidence.

In close connection with the development corpus-based resources for MWEs, the creation of gold standards, possibly manually controlled, proves essential for the evaluation of methods and resources, training of automated systems, and the development of portable language resources (domain specific, mono- and multilingual), whose compilation is extremely expensive and difficult to control.

There is now a large number of lexicons and dictionaries specifically dedicated to MWEs, fixed phrases and idioms in many languages and in multilingual environments. The question raised by MWEs within lexicon building is closely connected to their nature as linguistic items between lexis and syntax , and concerns deciding between, on the one hand, listing complex forms in the lexicon and, on the other hand, writing rules to derive them; however, especially in a multilingual context, this distinction does not always hold since it is often the case that a MWE in one language matches with a free form (single word or transparent expression) in another language, or the reverse. Another crucial aspect in the development of MWE-related resources concerns the distinction between the elements of a MWE representing its core, obligatory components and those that are used by the lexicographer only to provide the user with grammatically acceptable examples: in fact, lexical resources are often an incomplete source of information on this topic, also considering that, even by means of corpus crawling, a clear citation form does not always seem to be easily detected since variations are often as frequent as forms accepted as citation forms. (Bond et al., 2005). If we look at the problem of resource design and implementation from a potentially multilingual (or language-nonspecific) perspective, as Calzolari et al. (2002) point out, one of the essential requirements is to create a shared model suitable for different languages, as well as determine and represent the links (such as constraints, conditions etc.) among different entries within a monolingual lexicon to facilitate translation and, in addition, make the monolingual resource exploitable by itself as a look-up source (but also flexible for a variety of purposes). This is why standardisation initiatives play a central role in the creation of resources for computational lexicography.

3. Experiments on Italian Data

3.1. Data collection and creation of an electronic lexicon

The *De Mauro Paravia Dictionary of the Italian Language* is a portion of GRADIT (De Mauro, 2000), the largest dictionary for Italian, and is freely available.[1] Several MWEs are listed under each lemma, wherever relevant[2]. Each MWE is stored only once, and appears in the context of the lemma that was considered most appropriate by the lexicographers. For instance, the expression "a mente fredda" ("in cold blood" [lit.: "with a cold mind"]), is listed under "mente" (mind) and not under "fredda" (cold) or "a".

On the basis of the information available from the De Mauro-Paravia online dictionary, we created an electronic lexicon of MWEs encoded in XML which preserves most of the information that could be obtained from the dictionary itself, such as the indication of the main lemma (which represents a single entry under which multiple MWEs can be assigned), the grammatical category of the resulting expression (feminine/masculine noun, adverbial expression, (in)transitive verbal expression, etc., referred to as the expression type), the sense of the lemma which is involved in the MWE, a gloss in the form of a description of the MWE meaning, the contexts of use (technical or common usage, specific domains etc.), and one or more example sentences. All MWEs pertaining to the same lemma are kept together under a single entry provided for that lemma and receive different distinctive numbers. A snapshot for the lemma "forza" (force, power, strength) is given in Figure 1.

In Table 1 and Table 4 we report the distribution of expression types and contexts of use in the database, respectively. Both kinds of information were inherited from the dictionary.

3.2. The Corpus

The corpus used for our research is *ItWac*, a two billion word corpus of Italian. *ItWac* was entirely built from the Web through several stages of crawling and cleaning (aimed at privileging precision over recall) and covers a maximally broad spectrum of registers, text genres, and topics, and has tags for part-of-speech and lemma, both automatically obtained (Baroni et al., 2009). A contrastive evaluation of *ItWac* against the 300 million token Repubblica corpus showed that although noise rate is clearly smaller for the latter, *ItWac* has a richer variety of lexical types and a larger number of usage examples and registers (Baroni et al., 2004). These features are essential for our research on MWEs.

3.3. Acquiring frequency information

Each MWE in our electronic database was transformed in a query interpreted by the Corpus Query Processor (cqp) and run on *ItWac*. All MWEs were searched as fixed strings, without allowing for form variations, although *ItWac* does have tags for lemmas. The main reason behind this choice is that one typical feature of (at least some) MWEs is *fixedness* or limited variability, so that allowing for lemma-based searches would dramatically increase the risk of collecting false positives, i.e. expressions that do not really represent an instance of the MWE although they feature the same lexical items (e.g. "vedere rosso" [MWE] vs "vedere rossi" [phrase]). Insights coming from the present study will guide us towards a reasonable approach to search expansion in this sense, so as to limit flexibility to those types of MWEs which might allow for it. Each expression was matched within the context of one sentence; lemma and part-of-speech tags about each token in the sentence were also extracted.

We extracted a total of nearly 13 million examples across all MWEs, with the most frequent expression being the conjunction "se non" ("if not"), occurring nearly 130,000 times. Interestingly, out of the total of 13,782 expressions searched for, 2,203 returned zero matches. The distribution of zero matches over expression types and registers are provided in Table 1 and Table 4, respectively. As far as the grammatical category (expression type) is concerned, one can see that besides the 60% of unmatched phonosymbolic MWEs, most affected forms are nouns (ca. 19% of expressions were never found) and verbs (over 15% not found). We see two reasons for these figures, one intrinsic to the nature of these kinds of MWEs, and one due to a choice we adopted in the search settings. The intrinsic feature is that content words/expressions are, by definition, a wider class than function words/expressions, and given their stronger communicative power, they are more likely to be created for figurative usage and one-off occasions. This is not surprising as this reflects, in many respects, the distribution of the single words in the lexicon, where a very big, open set of lexical items has an extremely small closed set of functional words as counterpart. The procedural reason is that we did not include form variation in the search (see above), and this choice has surely a heavier impact on the retrieval of forms that in principle could undergo variations, such as nouns and verbs, than on that of forms that could not.

As for zero matches across registers, it is not surprising to observe that technical, regional, and obsolete expressions tend to be less represented, even in such a large and varied corpus such as *ItWac*. This is something that must be taken into account when including MWEs in lexical resources, since it might not make much sense to store a very large number of technical expressions in general purpose dictionaries. This becomes even more significant when we consider that the total number of data-extracted examples that pertain to the "common" register is three times higher than those characterised by technical-specialistic usage (nearly 9M vs. nearly 3M), whereas the latter are nearly double the former in the dictionary.

In fact, the frequency distribution of the MWEs occurrences in the corpus largely confirm what con be observed by the

[1] The dictionary has been accessible for online browsing at `http://old.demauroparavia.it/index.php` for years, but since a few weeks at the time of writing it appears to be no longer available.

[2] (De Mauro, 2000) provides three criteria to recognise MWEs, namely i) the existence of a semantic surplus making the meaning non-compositional, i.e. not computable from that of the simplex componenets; ii) a certain degree of lexical and syntactic frozenness or limited variability; iii) a certain degree of institutionalisation or a significant presence in domain-specific languages.

```
<entry id="45379" lemma="for|za" gram="sostantivo femminile">
 <mwe id="1">
  <expression type="locuzione avverbiale">a, di forza</expression>
  <description sense="0" use="comune">con la forza</description>
 </mwe>
 <mwe id="2">
  <expression type="locuzione preposizionale">a forza di</expression>
  <description sense="0" use="comune">per indicare unazione ripetuta insistentemente</description>
  <example>a f. di insistere hai ottenuto quello che volevi</example>
 </mwe>
 <mwe id="3">
  <expression type="locuzione avverbiale">a forza di braccia</expression>
  <description sense="0" use="comune">con grande sforzo, specialmente fisico</description>
 </mwe>
 <mwe id="4">
  <expression type="locuzione avverbiale">con la forza</expression>
  <description sense="0" use="comune">impiegando mezzi coercitivi; con violenza</description>
  <example>costringere con la f.</example>
 </mwe>
 <mwe id="5">
  <expression type="locuzione sostantivale femminile">forza a distanza</expression>
  <description sense="0" use="tecnico-specialistico"><term>fisica</term>, f. che si esercita tra
corpi posti a distanza in un mezzo materiale o nel vuoto</description>
 </mwe>
</entry>
```

Figure 1: Snapshot of the XML database of Italian MWEs.

Table 1: Expression types in database (DB), examples, zero (∅) matches in *ItWac*, percentage of ∅ matches over cases in DB, most frequent pattern observed per type, and its frequency

information from DB			**information from corpus data**			
			zero matches		**morpho-syn patterns**	
expression type	**example**	**# in DB**	**# ∅ matches**	**%**	**most freq pattern**	**freq**
masculine noun	“gas di scarico”	5,377	1,045	19.4	NOUN_ADJ	2,038
feminine noun	“adozione a distanza”	4,905	941	19.2	NOUN_ADJ	2,175
adverb	“seduta stante”	1,434	46	3.2	PRE_NOUN	516
adjective	“nudo e crudo”	828	23	2.8	PRE_NOUN	406
verbal form	“marcare a uomo”	827	129	15.6	VER:infi_ART_NOUN	229
preposition	“riguardo a”	196	2	1.0	PRE_NOUN_PRE	96
conjunction	“invece di”	90	2	2.2	ADV_CHE	18
interjection	“nemmeno morto”	72	3	4.1	ADV_PRE_NOUN	11
phonosymbolic expression	“zic zac”	8	5	62.5	–	–
pronoun	“alcun che”	7	0	–	PRO:pers_PRO:indef	2
indefinite pronoun	“chi capita”	3	0	–	–	–
relative indefinite pronoun	“che che”	1	0	–	CHE_CHE	1
unspecified	–	34	7	2.1	–	–
total	–	13,782	2,203	–	–	–

anaysis of zero matches. Table 2 reports the number of tokens (i.e. the positive hits returned by the crawling procedure) for every expression type, along with the percentage, already reported in Table 1, regarding the distribution of expression types in the lexicon, to facilitate comparisons. Table 3 shows the frequency distribution across different registers.

3.4. Creation of example corpora

For each target idiom that returned non-null matches, an example corpus was created containing all the relevant strings drawn form the crawling of ItWac. Structural information such as sentence delimiters was retained, as was the annotation for part-of-speech and lemmas. Every example corpus is stored as plain text in a separate directory.

A snapshot of an example corpus directory (for the expression “a conduzione familiare” (“family-run”)) along with a sample of KWIC concordances is given in Figure 2.

Table 3: Distribution of registers as from ItWac

register	**# in itWac**
technical-specialised	2,842,120
common	8,912,636
common, tech-spec.	398,235
very informal	250
only literary	142,834
obsolete	12,072
regional	5352
obsolete, only literary	38,195
obsolete, tech-spec.	5,256
very formal	—
total	12.721M

3.5. Acquiring morphosyntactic patterns

To investigate the internal structure of the MWEs, also towards the identification of relevant detection features,

Table 2: Distribution of expression types as from ItWac

expression type	# of tokens	% over all tokens	% over total Lexicon
adverbial	4,891,117	38.45%	10.4%
adjectival	2,539,864	19.96%	6%
nominal (feminine)	1,473,828	11.59%	35.6%
nominal (masculine)	1,392,822	10.95%	39%
conjunctional	1,028,722	8.09%	0.7%
prepositional	993,279	7.81%	1.4%
interjections	200,227	1.57%	0.5%
verbal	73,100	0.57%	6%
unspecified	63,381	0.50%	0.25%
pronominal	42,031	0.33%	0.06%
pronominal (relative indefinite)	20,122	0.15%	0.01%
pronominal (indefinite)	2,970	0.02%	0.02%
phonosymbolic	169	0.001%	0.06%
total	12,721M	100%	100%

1395550: [...] Hotel/NPR/Hotel Doge/NOUN/doge L'/ART/l' Hotel/NPR/Hotel Doge/NOUN/doge è/VER:fin/essere un/ART/un piccolo/ADJ/piccolo hotel/NOUN/hotel a/PRE/a tre/DET:num/tre stelle/NOUN/stella <**a/PRE/a conduzione/NOUN/conduzione familiare/ADJ/familiare**> gestito/VER:ppast/gestire da/PRE/da veneziani/NOUN/veneziano ./SENT/. [...]
5690956: [...] ./SENT/. locale/ADJ/locale <**a/PRE/a conduzione/NOUN/conduzione familiare/ADJ/familiare**> ./SENT/. [...]

Figure 2: Portion of example corpus for the MWE "a conduzione familiare" ("family-run")

Table 4: Registers in lexicon, zero (∅) matches in *ItWac*, and percentage of ∅ matches over cases in lexicon.

register	# in DB	# ∅ matches	%
technical-specialised	8,473	1,846	21.8
common	4,529	260	5.7
common, tech-spec.	300	15	5.0
very informal	250	39	15.6
only literary	105	5	4.8
obsolete	49	16	34.8
regional	40	16	40.0
obsolete, only literary	19	1	5.3
obsolete, tech-spec.	16	5	31.2
very formal	1	0	–
total	13,782	2,203	–

Table 5: Ten most frequent observed morpho-syntactic patterns for MWEs.

pattern	# MWEs	# occurrences
NOUN_ADJ	4,236	1,595,101
NOUN_PRE_NOUN	1,686	549,549
PRE_NOUN	933	4,139,954
NOUN_ARTPRE_NOUN	666	186,594
ADJ_NOUN	244	109,981
ARTPRE_NOUN	233	1,209,764
VER:infi_ART_NOUN	229	24148
NOUN_NOUN	191	73,935
ADJ_ADJ	130	11,935
NOUN_VER:ppast	126	50,259

we created morpho-syntactic patterns based on the part of speech information we collected in the extraction of examples. These were then matched with the information about expression types provided in the database. In Table 5 we report the 10 most frequent patterns, each characterising more than 100 MWEs, and covering in total 75% of the expressions in the database, together with two figures: the number of expressions that exhibit that pattern, and the overall number of occurrences (i.e. the sum of occurrences of all MWEs with that pattern).[3]
The last two columns in Table 1 show instead the most frequent observed pattern for each expression type and the number of MWEs of that type which exhibit it. No pattern is reported for types where there is extreme variation (all patterns encountered are different). Apart from the NOUN_NOUN structure, which characterises nearly exclusively expressions resulting in nominal forms, and is typical of some compounds in Italian which often indeed exhibit features of MWEs, no striking idiosyncratic pattern can be observed, at least in the most frequent patterns reported in this paper. Additionally, identical patterns are found for different resulting expression types (PRE_NOUN identifies both adverbial and adjectival forms, see Table 1). It is also interesting to note that the most frequent MWE ("se non") exhibits an idiosyncratic pattern (CON_NEG) which no other expression in the database appears to show. Finally, we must bear in mind that the observations based on the patterns are subject to errors in the automatic tagging of the corpus, and that due to the often morpho-syntactically idiosyncratic nature which characterises MWEs, the fre-

[3] Rarely, the same MWE occurring in different syntactic contexts was tagged differently, thereby yielding a different pattern. Thus, for each expression we only selected the most frequent exhibited pattern.

quency of POS tagging errors in this context might be higher than average.

3.6. Creation of a MWE database

The way the acquired information is stored and organised is a crucial issue. Although guidelines exist, the database design was not a trivial task, for often similar tools are devised for specific tasks and do not allow for the flexibility that we had in mind for our resource.

On the basis of the morphosyntactic patterns extracted from ItWac, we created a matrix that encodes such information and allows for specific queries, taking the final form of a relational database. Morphosyntactic patterns constitute the relations in the database, single parts-of-speech are represented as attributes (the columns of the matrix) while the MWEs are represented as tuples whose elements enter the appropriate relation according to their pos-tag (which is eliminated from the database as it would represent redundant information).

Steps in the design of the database included i) the automatic generation of the ordered string of part-of-speech tags, realised using simple php scripting. Clearly, the string must subsume all possible instances of the morphosyntactic patterns and at the same time maintain their original order, so that a string of NOUN_NOUN_ADJ subsumes NOUN_ADJ but does not subsume NOUN_ADJ_NOUN, for which specific constraints have been set; ii) mapping all the entries onto the original pos-string, and writing strings in the form of arrays generated by writing the relevant component in the appropriate position when the pos-tag of the entry component and the pos-tag of the original string matched, and a standard separator (;) otherwise; iii) manual check to correct mistakes produced both by the source data and during the DB implementation procedure.

The resulting product is very flexible: it is nothing more than a text document, but it encodes all the relevant information and can be easily modified at one own will (for example by substituting the separator with other symbols, by adding or replacing tags at the top of the matrix etc.) and can be opened with any spreadsheet editor. A snapshot of the resulting database obtained using Microsoft Excel is given in Figure 3. Filters can be applied to data so that queries over single items or cross researches can be made.

4. Discussion

From the figures reported in the experiment results, we notice that as for the frequency distribution of entries over the source lexicon, content MWEs represent the striking majority (with a strong preference for nominals) as opposed to the very limited variety of function MWE. This was expected as it reflects the usual distribution of simplex items over the lexicon.

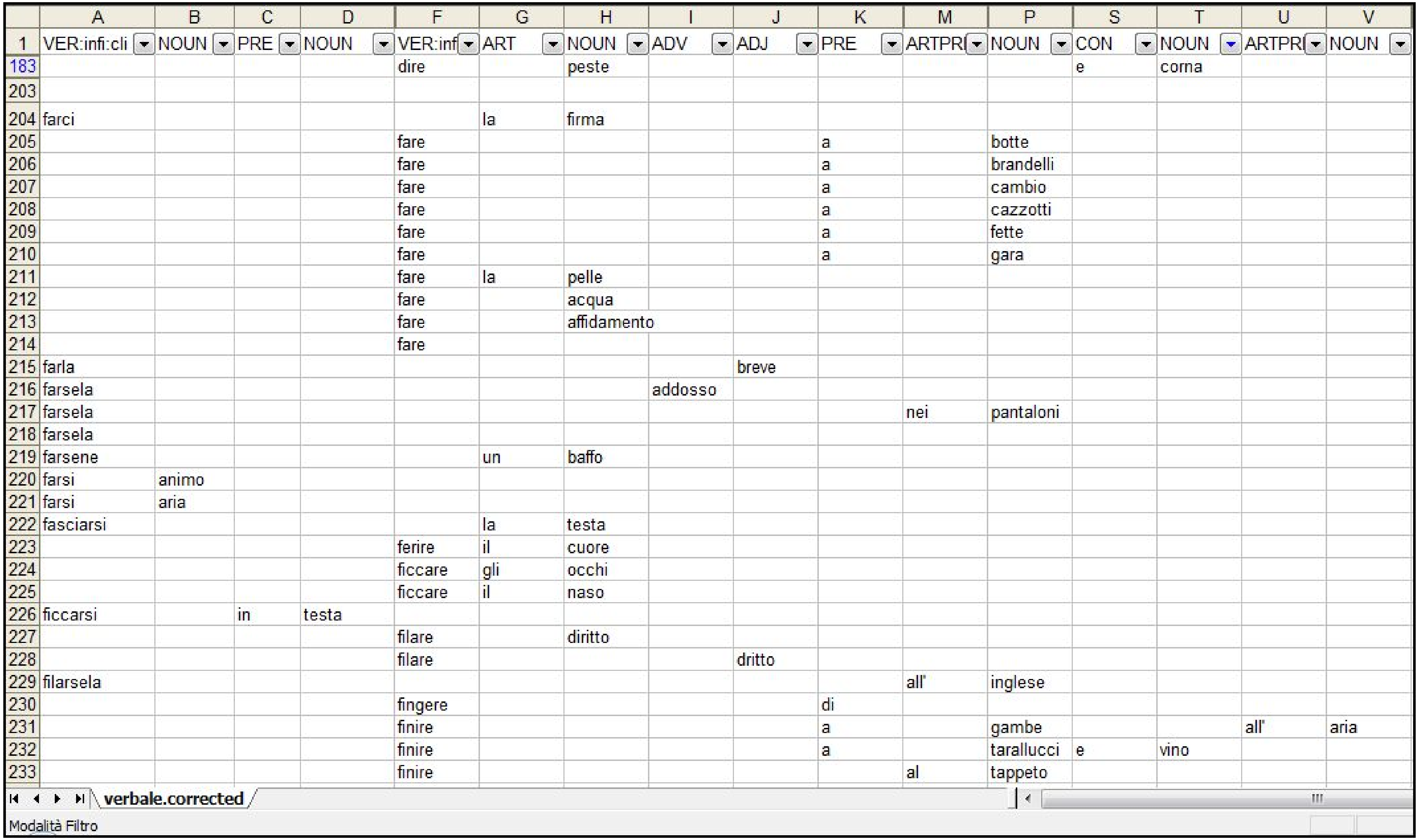
However, corpus data has shown that zero matches reverse some tendencies as the most affected content categories are nominals (around 39% altogether) and verbal expressions (15%). This may be a consequece of choices in the query procedure and of the intrinsic nature of these expressions, as discussed above. The number of zero matches is, among other things, an indicator of two facts: i) especially in the case of verbs, it indirectly indicates that more fine-grained criteria (coming from theoretical and descriptive interpretation of the phenomena) must be considered to extend queries to inflected or modified forms in order to gain more authentic insight on the phenomenon; ii) the idea of a 'quotation form' being the most frequent configuration in which an expression is found can harly be held for MWEs;

As for the distribution of registers, strong preference has been given by the lexicographers to technical-specialised domains (covering more than half of the data) and common domains (around 30%). Again, corpus matches show that technical, regional, and obsolete expressions tend to be less represented, even in such a large and varied corpus such as itWac. In particular, technical-specialised domains are the most affected, representing 38% of unmatched entries. This is something that must be taken into account when including MWEs in lexical resources, since it might not make much sense to store a very large number of technical (or regional) expressions in general purpose dictionaries.

The fact that data-extracted examples that pertain to the "common" register is three times higher than those characterised by technical-specialised usage (see Section 3.3.) confirms the idea that MWEs are far from being an 'appendix' of the language. In fact, data regarding the frequency distribution of tokens across categories and registers tend to confirm the observations and the generalisation drawn for zero matches. Adverbial types are the best represented, which is expected since they combine both the property of being lexical elements and the fact of being less biased by the choice of querying the entries as fixed strings, because they hardly ever allow for form variations. It is however surprising to see that adjectival expressions are less affected than other expressions in terms of occurrences: also looking at their internal morphosyntactic structure, it seems that they are much more similar to adverbial expressions than to other kinds of lexical MWEs (and therefore possess a greater deal of fixedness than nominals or verbal expressions). Most probably the boundaries between these two groups can hardly be drawn on the basis of structural evidence, while it in theory it should be possible to separate them only on the basis of distributional features. Prepositional expressions are very well represented in the corpus and also looking at their internal structure it seems that they represent the borderline case between lexical and grammatical items, both in terms of their frequency, which is partly due to their invariable form, and by the fact that the expressions feature a wealth of simplex lexical (such as verbs and nouns) among their constituents. Finally, the frequency distribution of tokens across patterns within one type tend to confirm the rankings of the most represented type of pattern within that category. When there is an extreme variety (as for interjective, phonosymbolic and pronominal expressions) the frequency distribution of tokens gives a clear characterisation of the most salient morphosyntactic configurations among these expression types.

5. Future developments

Although not all MWEs are necessarily collocations, in future research we plan to exploit standard association measures, such as the mutual information or the log-likelihood



	A	B	C	D	F	G	H	I	J	K	M	P	S	T	U	V
1	VER:infi:cli	NOUN	PRE	NOUN	VER:inf	ART	NOUN	ADV	ADJ	PRE	ARTPR	NOUN	CON	NOUN	ARTPR	NOUN
183					dire		peste						e	corna		
203																
204	farci					la	firma									
205					fare					a		botte				
206					fare					a		brandelli				
207					fare					a		cambio				
208					fare					a		cazzotti				
209					fare					a		fette				
210					fare					a		gara				
211					fare	la	pelle									
212					fare		acqua									
213					fare		affidamento									
214					fare											
215	farla								breve							
216	farsela							addosso								
217	farsela										nei	pantaloni				
218	farsela															
219	farsene					un	baffo									
220	farsi	animo														
221	farsi	aria														
222	fasciarsi					la	testa									
223					ferire	il	cuore									
224					ficcare	gli	occhi									
225					ficcare	il	naso									
226	ficcarsi		in	testa												
227					filare		diritto									
228					filare				dritto							
229	filarsela										all'	inglese				
230					fingere					di						
231					finire					a		gambe			all'	aria
232					finire					a		tarallucci	e	vino		
233					finire						al	tappeto				

verbale.corrected

Modalità Filtro

Figure 3: Snapshot of database of morphosyntactic patterns for the verbal expressions

ratio to obtain an index of cohesion within the expression which can be possibly combined with raw frequency to determine the direction of their further treatment.

We will also relax the "fixedness" constraint imposed on this preliminary search via including lemmatised forms only for those type of expressions which are more likely to allow for morpho-syntactic variation. Overall, for various reasons, morphosyntactic information might prove useful in detecting and classifying MWEs probably only if combined with indicators of different nature, such as for instance lexico-semantic features.

Another avenue to explore towards MWE extraction is a generalisation at the semantic level. The morphosyntactic matrix shows that some MWEs do allow for a quite high degree of lexical flexibility. For instance, the verbs that take "head" ("la testa") as an object are all synonyms ("abbassare", "chinare", "curvare"). Similarly, the direct object of the verb "abbassare" is often a body part. In the first case, it seems that the use of a given verb is not strictly constrained, and a synonym can take its place. In the second case, one can observe a generalisation in terms of object type, in the same fashion as selectional preferences (all objects are cohyponyms). In both cases, exploiting an existing lexical resource such as an Italian WordNet (for instance the freely available MultiWordNet) could be an automatic way of obtaining abstractions over semantic classes (such as "body part" for "testa", "tail", "arm") through hypernym links, and synonyms (such as "chinare" for "abbassare") thanks to synsets. This might allow for guided extraction of MWEs which are not yet in the database. As for a longer-term aim, we hope to gather insights that will allow us to model the form and use of valid MWEs towards a general characterisation of these expressions and their automatic detection. Given the nature of MWEs, generalisations of this kind are likely to apply only to a small portion of MWEs (those featuring a minimal degree of fixedness), so that overgeneration of expressions is an actual risk. However, the generation of invalid expressions is only theoretical, since all of the potential MWEs are to be checked against a corpus and only the occurring phrases will be retained. This should allow for an expansion of the database in a structured manner, still privileging precision over recall.

6. 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*.

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 49:209–226.

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.

A. Copestake, F. Lambeau, A. Villavicencio, F. Bond, T. Baldwin, I. A. Sag, and Dan Flickinger. 2002. Multiword expressions: Linguistic precision and reusability. In *In Proceedings of LREC 2002*, Las Palmas, Spain.

CSL. 2005. *Special issue on Multiword Expressions of Computer Speech & Language*, volume 19.

T. De Mauro. 2000. *GRADIT, Grande dizionario italiano dell'uso*. UTET.

C. Fellbaum, editor. 2007. *Collocations and Idioms: Corpus-Based Linguistic and Lexicographic Studies*. Continuum Press, Birmingham.

JLRE. 2009. *Special issue on Multiword Expressions of the Journal of Language Resources and Evaluation*, volume to appear.

R. Moon. 1998. *Fixed Expressions and Idioms in English*. Clarendon Press, Oxford.

G. Nunberg, I. Sag, and T. Wasow. 1994. Idioms. *Language*, 70(3).

S.L. Piao, G. Sun, P. Rayson, and Q. Yuan. 2006. Automatic extraction of chinese multiword expressions with a statistical tool. In *Proceedings of the EACL Workshop on Multi-word expressions in a Multilingual Context*, pages 17–24.

P. Rayson, D. Archer, S. Piao, and T. McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the LREC 2004 Workshop, beyond Named Entity Recognition Semantic Labelling for NLP Tasks*, pages 7–12.

P. Rayson, S. Piao, S. Sharoff, S. Evert, and B. Villada Moirón. 2009. Multiword expressions: hard going or plain sailing? *Journal of Language Resources and Evaluation*. To appear.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing: Third International Conference (CICLing 2002)*, Berlin/Heidelberg. Springer.

A. Villavicencio, A. Copestake, B. Waldron, and F. Lambeau. 2004. Lexical encoding of MWEs. In *Proceedings of the ACL'04 Workshop on Multiword Expressions*, pages 80–87.

Acknowledgments

We are grateful to Marco Baroni and Eros Zanchetta for letting us use a server equipped with *ItWac* and cqp where we could run our experiments. We would also like to thank the anonimous reviewers for their useful comments and insight.