POS-PATTERNS OR SYNTAX? COMPARING METHODS FOR EXTRACTING WORD COMBINATIONS

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

Gianluca E. Lebani
University of Pisa
gianluca.lebani@for.unipi.it

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

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

Malvina Nissim
University of Groningen
m.nissim@rug.nl

Lucia Passaro
University of Pisa
lucia.passaro@for.unipi.it

Abstract

This paper reports on work carried out in the framework of an ongoing project aimed at building an online, corpus-based lexicographic resource for Italian Word Combinations. Our aim is to compare two of the most commonly used methods for the automatic extraction of word combinations from corpora, with a view to evaluate their performance – and ultimately their efficacy – with respect to the task of acquiring word combinations for inclusion in the lexicographic combinatory resource.

1. WORD COMBINATIONS: LEXICOGRAPHY AND NLP

It is widely acknowledged that lexicographers' introspection alone cannot provide comprehensive information about word meaning and usage, and that investigation of language in use is fundamental for any reliable lexicographic work (Atkins and Rundell 2008). This is even more true for dictionaries that record the combinatorial behaviour of words, where the lexicographic task is to detect the typical combinations a word participates in. In fact, it was much harder to study lexical combinatorics empirically before the advent of large corpora and the definition of statistical techniques for the analysis of word associations (Hanks 2012).

This paper reports on work carried out in the framework of an ongoing project called CombiNet[31] aimed at building an online, corpus-based lexicographic resource for Italian Word Combinations. We use the term **Word Combinations** (WoCs) to encompass both Multiword Expressions (MWEs) – namely WoCs characterised by different degrees of fixedness and idiomaticity that act as a single unit at some level of linguistic analysis, such as idioms, phrasal lexemes, collocations, preferred combinations (Calzolari et al. 2002, Sag et al.

[31] **PRIN Project 2010-2011** *Word Combinations in Italian* (n. 20105B3HE8) funded by the Italian Ministry of Education, University and Research (MIUR). URL: http://combinet.humnet.unipi.it.

2002, Gries 2008) – and the distributional properties of a word at a more abstract level (argument structure, subcategorization frames, selectional preferences), along the lines of Benson et al. (2010).

The specific aim of this paper is to compare two of the most commonly used methods for the automatic extraction of WoCs from corpora (cf. 1.1), with a view to evaluate their performance – and ultimately their efficacy – with respect to the task of acquiring WoCs for inclusion in our lexicographic combinatory resource. More specifically, we calculate the recall of the two methods using as benchmark the list of combinations recorded in *Dizionario Combinatorio Italiano* (DiCI, Lo Cascio 2013), the largest existing Italian combinatory dictionary. In addition, manual inspection of the top candidates in both datasets is used to assess the proportion of *valid* WoCs that are extracted from the corpus but unattested in DiCI.

1.1. Comparing methods for WoC extraction

Currently, apart from purely statistical approaches, the most common methods for the extraction of WoCs involve searching a corpus via sets of patterns and then ranking the extracted candidates according to various association measures (hybrid method) in order to distinguish meaningful combinations from sequences of words that do not form any kind of relevant unit (Villavicencio et al. 2007, Ramisch et al. 2010).

Generally, the search is performed for either shallow morphosyntactic (POS) patterns (**P-based approach**) or syntactic dependency relations (**S-based approach**). In the case of P-based methods, one needs to have a POS-tagged corpus and to draw a list of POS-patterns assumed to be representative of WoCs in a given language (see e.g. (1)). In the case of S-based methods, one needs to have a parsed corpus and to identify syntactic relations that may give rise to meaningful WoCs (see e.g. (2)).

(1) a. NOUN PREP NOUN

punto di vista

'point of view'

b. NOUN ADJ

anno accademico

'academic year'

(2) a. SUBJ – VERB

guerra – scoppiare

'war – burst'

b. VERB – OBJ

perdere – vista

'lose – (one's)sight'

Most studies so far have concentrated on P-based approaches, which yield satisfactory results for relatively fixed, short and adjacent WoCs. More recently it has been suggested that syntactic dependencies might be helpful to also capture discontinuous and syntactically flexible WoCs, because they can extract syntactically related words irrespective of their surface realizations (Seretan 2011).

Clearly, both methods have cons. In the case of P-based methods, POS-patterns need to be specified a priori. Moreover, not every extracted combination is a WoC, even using a detailed list of patterns and even after applying association measures (cf., among others, Nissim et al. 2014). Finally, without considering syntactic information, it is difficult to extract complex and flexible WoCs (especially verbal ones), let alone more schematic combinatory information (e.g. argument structure). As for S-based methods, abstracting away from specific constructs and information (e.g. linear order, morphosyntactic features, etc.) may result in little information about how exactly words combine. Moreover, it is hard to distinguish frequent, regular combinations from highly fixed, idiomatic ones with the very same syntactic structure.

Overall, the two methods seem to be highly complementary rather than competing with one another. In fact, various attempts are currently being proposed to put them together (cf. the SYMPAThy method discussed in Lenci et al. 2014, 2015; cf. also Heid 2015 and Squillante 2015), and the results of our experiment also point in this direction.

2. THE EXPERIMENT

In order to test and compare the performance of the two above-mentioned methods with respect to the task of extracting WoCs for lexicographic purposes, we selected a sample of 25 Italian target lemmas (TLs) – including 10 nouns, 10 verbs and 5 adjectives (listed in Table 1) – and we extracted P-based and S-based combinatory information from *la Repubblica* corpus[32] (Baroni et al. 2004). TLs were selected by combining frequency information derived from the *la Repubblica* corpus and inclusion in Lo Cascio's (2013) DiCI, which is used for (part of the) evaluation.

Nouns	**Verbs**	**Adjectives**
anno 'year'	*parlare* 'talk / speak'	*economico* 'economic'
governo 'government'	*prendere* 'take'	*giovane* 'young'
casa 'house'	*tenere* 'keep / hold'	*basso* 'low / short'
fine 'end / goal'	*vivere* 'live'	*facile* 'easy'
guerra 'war'	*perdere* 'lose/miss'	*rosso* 'red'
famiglia 'family'	*uscire* 'go out'	
mano 'hand'	*lavorare* 'work'	
situazione 'situation'	*costruire* 'build'	
morte 'death'	*pagare* 'pay'	
stagione 'season'	*leggere* 'read'	

Table 1. Target lemmas for the experiment

[32] The *la Repubblica* corpus (approx. 380M tokens) contains texts from the homonymous Italian daily newspaper. The version of the corpus we used was POS-tagged with the tool described in Dell'Orletta (2009) and dependency-parsed with DeSR (Attardi and Dell'Orletta, 2009).

As regards the P-based method, we extracted all the occurrences of each TL in a set of 122 pre-defined POS-patterns deemed representative of Italian WoCs. The set includes:

- POS sequences mentioned in existing combinatory dictionaries (previously collected in Piunno et al. 2013) and relevant theoretical literature (e.g. Voghera 2004; Masini 2012);

- "new" patterns identified through corpus-based, statistical experiments (Nissim et al. 2014);

- more patterns added manually by elaborating on the previous lists.

For the actual extraction, we used the EXTra tool (Passaro & Lenci 2015). EXTra retrieves all occurrences of the specified patterns, only as linear and contiguous sequences (no optional slots can be included), and ranks them according to a variety of association measures, among which we chose Log Likelihood (LL). The search considers lemmas, not wordforms. Finally, only sequences with frequency >5 have been considered. See Table 2 for an example of data extracted with EXTra.

LL	FREQ	W1	POS	W2	POS	W3	POS
5176.86	702	medico	s	**di**	e	famiglia	s
5176.86	100	medico	s	**in**	e	famiglia	s
5176.86	9	medico	s	**di**	ea	famiglia	s
3205.18	6	amico	s	e	cc	famiglia	s
3205.18	82	amico	s	di	ea	famiglia	s
3205.18	545	amico	s	di	e	famiglia	s
2983.87	403	famiglia	s	cristiano	a		
2615.41	80	cassaforte	s	di	ea	famiglia	s
2615.41	152	cassaforte	s	di	e	famiglia	s
2537.23	600	intero	a	famiglia	s		
2315.64	1154	grande	a	famiglia	s		
2114.56	234	gioiello	s	di	e	famiglia	s

Table 2. Examples of candidates extracted by EXTra for the TL *famiglia* 'family'

Note that the same combination of lemmas can be listed more than once in the results: take for instance the lemma sequence *medico+di+famiglia* ('doctor'+'of'+'family'), which appears in row 1 and row 3 in Table 2. The two hits represent two separate candidates because of the different morphosyntactic configurations of the combination, i.e. because of the different preposition intervening between 'doctor' and 'family': a simple preposition in the first case (cf. *medico di famiglia* 'general practitioner, GP', which is indeed a MWE), an articulated preposition in the second case (cf. *medico della famiglia* 'doctor of the family', which is a normal phrase). Although the two candidates have the same LL value (5176,86), because the preposition is ignored when computing the association strength, their respective frequency (702 vs. 9) appears indicative. Moreover, since data extraction is based on shallow sequences, word order is strictly preserved in the output: hence, the combination *intero+famiglia* ('entire'+'family') represents only the occurrences of the two lemmas in this order (*intera famiglia* 'entire family', A+N), despite the reversed one (*famiglia intera*, N+A) would also be possible.

As regards extraction based on syntactic dependencies (S-based), we extracted the distributional profile of each TL using the LexIt tool (Lenci et al. 2012), which works with Italian nouns, verbs and adjectives. The LexIt distributional profiles contain the syntactic slots (subject, complements, modifiers, etc.) and the combinations of slots (frames) with which words co-occur, abstracted away from their surface morphosyntactic patterns and actual word order. For instance, *Gianni ha dato volentieri un libro a Maria* 'John has willingly given a book to Mary' and *Gianni ha dato a Maria un libro* 'John has given Mary a book' are both mapped onto the syntactic frame "subj#obj#comp_a", despite the different order of their slots and the presence of adverbial modifiers. Moreover, each slot is associated with lexical sets formed by its most prototypical fillers. The statistical salience of each element in the distributional profile is estimated with LL.

For each TL we extracted all its occurrences in different syntactic frames together with the lexical fillers (lemmas) of the relevant syntactic slots, abstracting away from their surface morphosyntactic patterns. As for the P-based settings, only combinations with frequency >5 have been considered.

LL	FREQ	W1 (POS)	SYNT_REL	W2 (POS)
8939.28	1258	famiglia (s)	**modadj-post**	reale (a)
7084.59	1577	famiglia (s)	**modadj-pre**	grande (a)
6364.01	1657	famiglia (s)	modadj-post	italiano (a)
4543.05	719	famiglia (s)	modadj-pre	intero (a)
4271.25	548	famiglia (s)	modadj-post	cristiano (a)
3740.05	514	famiglia (s)	modadj-post	mafioso (a)
3708.22	465	famiglia (s)	**comp_di**	vittima (s)
LL	FREQ	W1 (POS)	SYNT_REL	W2 (POS)
15128.3	1180	perdere (v)	comp_di	vista (s)
15118.06	2615	perdere (v)	**obj**	occasione (s)
12066.27	3539	perdere (v)	**obj**	tempo (s)
11360.72	1831	perdere (v)	obj	terreno (s)
6504.6	1475	perdere (v)	obj	testa (s)

Table 3. Examples of candidates extracted by LexIt for the TLs *famiglia* 'family' and *perdere* 'to lose'

Table 3 shows an example of data extracted with LexIt. Although word order is generally underspecified, in some cases it is indicated in the syntactic relation itself: for instance, the "modadj-post" relation indicates that the first candidate – composed of *famiglia*+*reale* 'family'+'royal' – is *famiglia reale* 'royal family' in the N+A order, whereas "modadj-pre" indicates that the second candidate – composed of *famiglia*+*grande* 'family'+'big' – is *grande famiglia* ('big family') as A+N. Also, note that, in LexIt frames intervening tokens between slots (e.g. determiners, adverbial modifiers, etc.) are not recorded. Hence, the difference between *perdere*+*occasione* 'miss'+'chance' (which normally requires a determiner: *perdere un'occasione* 'miss a chance') and *perdere*+*tempo* 'lose'+'time' (where *tempo* is typically a bare noun: *perdere tempo* 'waste time') is not captured.

3. EVALUATION

The performance of the two extraction methods was assessed by means of a twofold evaluation. First, we calculated precision and recall using as benchmark dataset the list of combinations for our TLs recorded in DiCI. This was expected to shed light on the independent performance of the two methods overall, and with respect to the extraction of different types of WoCs. In addition, human evaluation of the top P-based and S-based candidates was carried out to assess the proportion of valid WoCs that are extracted from the corpus but unattested in a manually compiled resource like DiCI, thus providing information towards improving dictionary coverage.

3.1. Evaluation against DiCi

As DiCI is a traditional paper dictionary, we first built our gold standard benchmark dataset by digitizing the relevant entries and stripping off irrelevant information to obtain bare WoC lists. In order to enable automatic comparison with candidates from the two extraction systems, we then obtained a lemmatized version of benchmark combinations by performing POS and lemma annotation with the same tools used for corpus processing. Then we calculated global recall, overall precision and R-precision, as discussed in the following sections.

3.1.1. Recall

Recall is calculated as the percentage of extracted candidates out of the combinations found in the gold standard. For example, for the TL *rosso* 'red', EXTra extracts 23 of the 32 entries included in DiCi, thus its recall is 71.9% (23/32).

Recall points to an overall complementarity of the two systems, which are biased towards targets with different POS. As shown by the dark grey cells in Table 4, apart from cases in which the two systems have a very close performance (light grey cells), EXTra (P-based) performs better than Lexit (S-based) for nominal and adjectival TLs, whereas LexIt has a higher recall for virtually all verbal TLs.

Lemma		DiCi	EXTra_cand	Over	Rec	LexIt_cand	Over	Rec
rosso	A	32	805	23	0,719	476	22	0,688
situazione	N	149	2518	96	0,644	1343	79	0,53
stagione	N	64	1020	41	0,641	644	32	0,5
governo	N	144	5826	90	0,625	1327	50	0,347
anno	N	113	8762	63	0,558	3223	38	0,336
famiglia	N	130	2340	69	0,531	716	28	0,215
morte	N	83	1403	43	0,518	510	15	0,181
casa	N	356	3734	170	0,478	1092	95	0,267
mano	N	252	2555	117	0,464	934	34	0,135
facile	A	36	876	16	0,444	549	10	0,278
fine	N	71	2801	26	0,366	1017	11	0,155
economico	A	84	2384	62	0,738	981	62	0,738
guerra	N	62	2480	45	0,726	899	44	0,71
perdere	V	145	1557	96	0,662	2437	92	0,634
basso	A	72	668	46	0,639	457	44	0,611
giovane	A	50	1566	20	0,4	926	23	0,46
uscire	V	116	2010	66	0,569	2749	72	0,621
pagare	V	120	1474	61	0,508	1786	76	0,633
prendere	V	237	2831	109	0,46	4813	140	0,591
lavorare	V	98	1553	45	0,459	2218	53	0,541
vivere	V	197	1717	86	0,437	2517	106	0,538
leggere	V	117	1091	50	0,427	1514	68	0,581
costruire	V	90	1095	36	0,4	1249	53	0,589
parlare	V	194	3813	73	0,376	5896	87	0,448
tenere	V	159	1859	58	0,365	3569	97	0,61

Table 4. Comparing Extra and LexIt: Recall

3.1.2. Precision

Overall precision is not very significant as the number of extracted candidates for the two systems varies a lot, and it is generally very high. A better indicator of precision is *R-precision*, a measure borrowed from information retrieval and useful when assessing the quality of ranks. R-precision measures precision at the rank position corresponding to the number of combinations found in the gold standard, in our case DiCI. The rationale behind this is that an optimal system would place in the top *n* hits exactly all *n* entries found in the gold standard. Because our entries are ranked via association measures, and because both systems extract a large number of candidates, R-precision is a useful indicator of how well both methods perform and compare. To give an example, for the TL *pagare* 'to pay' there are 120 WoCs in the benchmark dictionary. In the top 120 candidates for Extra and LexIt we find 22 and 38 of these WoCs, respectively. So R-precision is higher for LexIt. Indeed, R-precision is almost always higher for LexIt (S-based) than for Extra (P-based), irrespective of POS, since Lexit performs better for all verbs and adjectives, as well as for most nouns (see dark grey cells in Table 5).

Lemma		total_DiCi	Extra	R-Prec	%	Over	Prec	Lexit	R-Prec	%	Over	Prec
pagare	V	120	1474	22	0,183	61	**0,041**	1786	38	**0,317**	76	0,034
tenere	V	159	1859	30	0,189	58	**0,031**	3569	41	**0,258**	97	0,016
perdere	V	145	1557	32	0,221	96	**0,062**	2437	36	**0,248**	92	0,039
costruire	V	90	1095	11	0,122	36	**0,033**	1249	21	**0,233**	53	0,029
vivere	V	197	1717	25	0,127	86	**0,05**	2517	43	**0,218**	106	0,034
prendere	V	237	2831	40	0,169	109	**0,039**	4813	50	**0,211**	140	0,023
uscire	V	116	2010	21	0,181	66	**0,033**	2749	22	**0,19**	72	0,024
leggere	V	117	1091	14	0,12	50	**0,046**	1514	21	**0,179**	68	0,033
lavorare	V	98	1553	16	0,163	45	**0,029**	2218	17	**0,173**	53	0,02
parlare	V	194	3813	20	0,103	73	**0,019**	5896	29	**0,149**	87	0,012
economico	A	84	2384	7	0,083	62	0,026	981	28	**0,333**	62	**0,063**
basso	A	72	668	13	0,181	46	0,069	457	18	**0,25**	44	**0,101**
rosso	A	32	805	2	0,062	23	0,029	476	6	**0,188**	22	**0,048**
giovane	A	50	1566	4	0,08	20	0,013	926	6	**0,12**	23	**0,022**
facile	A	36	876	2	0,056	16	0,018	549	4	**0,111**	10	**0,029**
situazione	N	149	2518	21	0,141	96	0,038	1343	40	**0,268**	79	**0,071**
guerra	N	62	2480	1	0,016	45	0,018	899	16	**0,258**	44	**0,05**
stagione	N	64	1020	8	0,125	41	0,04	644	16	**0,25**	32	**0,064**
casa	N	356	3734	43	0,121	170	0,046	1092	73	**0,205**	95	**0,156**
governo	N	144	5826	11	0,076	90	0,015	1327	22	**0,153**	50	**0,068**
famiglia	N	130	2340	17	0,131	69	0,029	716	19	**0,146**	28	**0,096**
anno	N	113	8762	2	0,018	63	0,007	3223	16	**0,142**	38	**0,02**
morte	N	83	1403	13	**0,157**	43	0,031	510	10	0,12	15	**0,084**
mano	N	252	2555	57	**0,226**	117	0,046	934	25	0,099	34	**0,125**
fine	N	71	2801	4	**0,056**	26	0,009	1017	3	0,042	11	**0,026**

Table 5. Comparing Extra and LexIt: R-precision

3.1.3. Thresholds

Obviously, precision and recall vary as we examine more candidates. This sort of information is useful when automatically extracted data then need to be analyzed manually by lexicographers. We therefore calculated both precision and recall at different thresholds (viz. every 250 hits). Figure 1 shows how they vary with increasing batch sizes; figures are averaged across different TLs with the same POS, so that we have one curve for nouns, one for adjectives, and one for verbs.

As expected, recall increases and precision decreases for both EXTra and LexIt . However, some interesting remarks can also be made. For example, apart from a few isolated cases which are represented by only few data points, recall for EXTra (top left) for nominal and verbal TLs seems to plateau after about 2,000 hits: this might suggest that a lexicographer could obtain a good coverage by concentrating on the manual evaluation of about 2,000 candidates per such TLs.

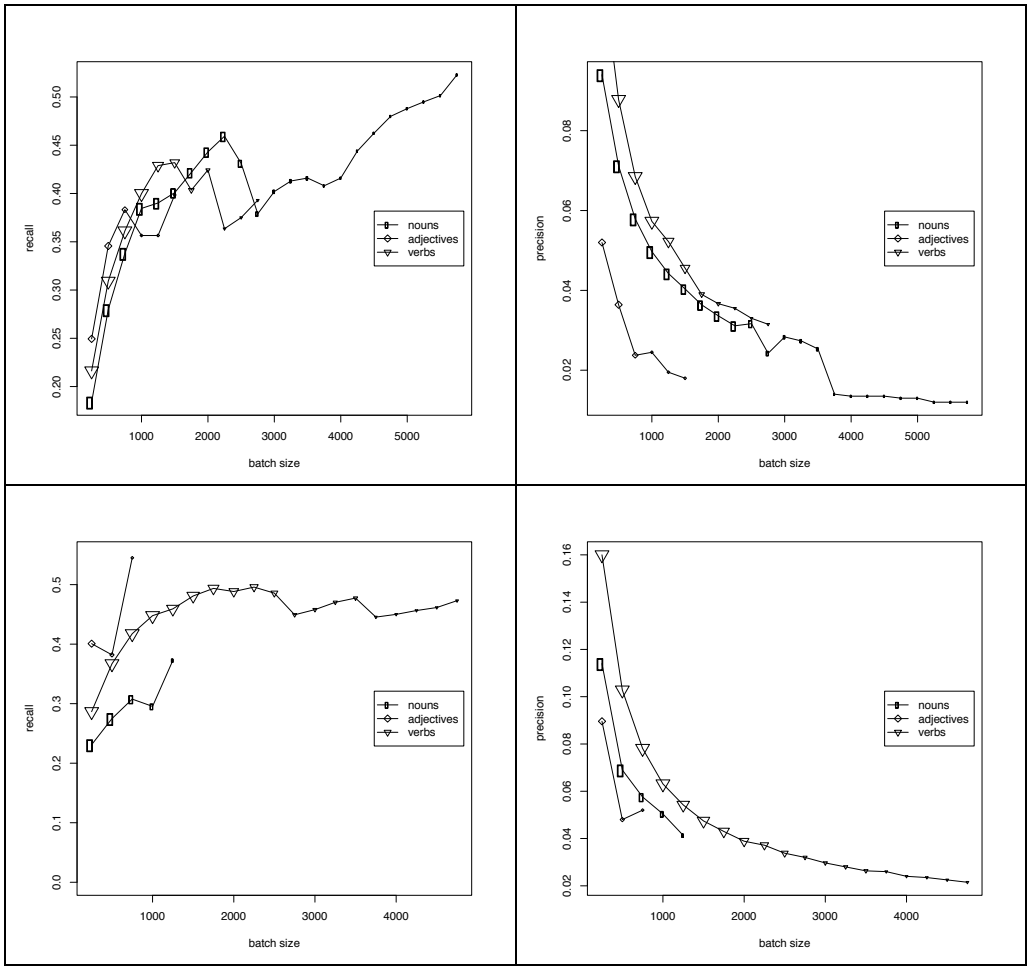

Figure 1. Precision and recall for EXTra (top) and LexIt (bottom)
plotted against batch sizes. The size of the data points indicates the number of
TLs included in the counts. The maximum size is 10 for nouns and verbs, and 5 for adjectives.
The minimum number is 2. Batches with only one TL are not shown.

3.1.4. P-based/S-based overlap

Total overlap is calculated as the percentage of cases in which EXTra and LexIt retrieve/don't retrieve the same gold standard combinations. For instance, the benchmark entry for the adjectival lemma *giovane* 'young' contains 50 combinations: out of these, 20 are retrieved by both EXTra and LexIt, 27 are retrieved by neither of the two systems, and LexIt only extracts 3 further WoCs. This means that the performance of the two systems is identical for 94% of DiCI combinations for the TL. This is the case of the highest overlap between the P-based and the S-based system, which is however quite high (76.05% on average, spanning between 59.07% and 94%, see Table 6). Random manual observations were made to explore possible causes for cases of "negative overlap", that is gold standard combinations that neither of the systems extracts. On the one hand, these appear to include e.g. WoCs with corpus frequency ≤5, as well as proverbs/idioms, thus pointing to a possible impact of corpus type and size. On the other hand, the actual WoC-ness/representativeness of some combinations in the gold standard is somewhat debatable.

When there is no overlap, i.e. when the two systems extract different gold standard combinations, the data (see Table 6) confirm that:

- the independent contribution of EXTra is higher (grey cells) for nouns and most adjectives, and in many such cases LexIt's contribution is minimal (often less than 1/5 of the number of "new" WoC retrieved by EXTra).

- the independent contribution of LexIt is higher for virtually all verbs, and in most such cases the contribution of EXTra is less than half the independent contribution of LexIt.

Lemma		total_DiCI	both	EXTra_only	LexIt_only	none	%overlap
casa	N	356	81	**89**	14	172	71,07
mano	N	252	29	**88**	5	130	63,10
governo	N	144	46	**44**	4	50	66,67
famiglia	N	130	26	**43**	2	59	65,38
anno	N	113	34	**29**	4	46	70,80
morte	N	83	14	**29**	1	39	63,86
situazione	N	149	68	**28**	11	42	73,83
perdere	V	145	72	**24**	20	29	69,66
fine	N	71	11	**15**	0	45	78,87
stagione	N	64	30	**11**	2	21	79,69
facile	A	36	9	**7**	1	19	77,78
basso	A	72	40	**6**	4	22	86,11
guerra	N	62	40	**5**	4	13	85,48
rosso	A	32	20	**3**	2	7	84,38
economico	A	84	57	5	5	17	88,10
prendere	V	237	76	33	**64**	64	59,07
vivere	V	197	70	16	**36**	75	73,60
parlare	V	194	60	13	**27**	94	79,38
leggere	V	117	43	7	**25**	42	72,65
pagare	V	120	56	5	**20**	39	79,17
lavorare	V	98	40	5	**13**	40	81,63
uscire	V	116	61	5	**11**	39	86,21
costruire	V	90	34	2	**19**	35	76,67
tenere	V	159	57	1	**40**	61	74,21
giovane	A	50	20	0	**3**	27	94,00

Table 6. Comparing EXTra and LexIt: overlap and differences in WoC extraction

A quick look at the different combinations extracted by the two systems suggests that the results might be influenced by the specific features and settings of the tools. For instance, verbs ending in *-si* (e.g. *prender-**si** un raffreddore* 'catch a cold') are not captured by EXTra because they are not lemmatized as such in the corpus, while LexIt extracts them thanks to dedicated frames (e.g. subj#si#obj). EXTra also does not capture complex complements as in *prendere con le mani nel sacco* 'catch (s.one) red-handed', as long, complex

patterns like V+PREP+DET+N+PREP+N were not included in the POS pattern set used for candidate extraction, due to their great variability and questionable productivity. Moreover, the possibility to include optional slots in LexIt, contrary to EXTra's fixed POS patterns, might favor the better performance by the former with verbal TLs. The picture is different for nominal/adjectival TLs, where variation and flexibility are less marked than with verbs.

Further investigation is needed to assess the exact impact of these features and settings on the results. Some problems may be solved by varing the extraction parameters, while others directly relate to intrinsic limits to either P-based or S-based approaches.

3.2. Human evaluation

Manual inspection of the top candidates in both datasets was used to assess the proportion of *valid* WoCs that were extracted from the corpus but unattested in DiCI. We obtained human judgments over 2,000 candidates for 10 TLs (1,000 from EXTra and 1,000 from LexIt, taking the top 100 results for each TL from each system).

Nouns	**Verbs**	**Adjectives**
guerra 'war'	*prendere* 'take'	*basso* 'low / short'
famiglia 'family'	*tenere* 'keep / hold'	*rosso* 'red'
mano 'hand'	*uscire* 'go out'	
stagione 'season'	*pagare* 'pay'	

Table 7. Target lemmas used for human evaluation

Annotators were linguists, not necessarily working on WoCs, mainly with a background in translation and/or corpus work. We collected two judgments per candidate. Possible annotations included: **Y** (yes, this is a valid WoC), **N** (no, this is not a valid WoC) and **U** (uncertain, not sure/this may be *part* of a valid WoC). We considered as *valid* candidates only those which received either YY or YU. Table 8 summarizes the results:

	Valid candidates extracted from corpus	**Valid candidates not recorded in DiCI**
EXTra	408 (/1000)	273 (/408)
LexIt	447 (/1000)	261 (/447)
EXTra+LexIt	**855** (/2000)	**534** (/855)

Table 8. Results of human evaluation

Out of 2,000 total candidates, we obtained positive evaluations for 855 combinations (408 from EXTra, 447 from LexIt). Out of these 855 WoCs deemed valid by the annotators, 534 are not recorded in DiCI: 273 from EXTra, 261 from LexIt. If we intersect the two sets, we find that only 80 WoCs are in common, which means we have **454** actual *new* WoCs, which are retrieved thanks to the two corpus-based methodologies. This again confirms their complementary contribution to WoC mining.

4. DISCUSSION AND CONCLUSION

The goal of this paper was to compare two commonly used methods for the automatic extraction of WoCs from corpora – the P-based method and the S-based method – with a view to evaluate their performance and efficacy. To this aim, we set up a twofold evaluation of candidates extracted by two systems – EXTra and LexIt – implementing the two approaches.

As for automatic evaluation (cf. 3.1), recall against DiCi is good for both EXTra (P-based) and LexIt (S-based). In addition, the data suggest a **complementarity** of the two systems, as recall appears to be related to the POS of the TL: EXTra performs better than LexIt for nominal and adjectival TLs, whereas LexIt has a higher recall for virtually all verbal TLs. However, further investigations might be needed to ascertain the extent to which the results are influenced by corpus type, by the specific features and settings of the extraction tools, as well as by the quality of the gold standard.

As for human evaluation (cf. 3.2), our experiment shows that over 40% of WoCs extracted by EXTra and LexIt are deemed valid by human annotators, and that more than half of these valid candidates are not attested in DiCI. This result is even more remarkable if we consider that we only evaluated the top 100 candidates for each TL/system. Automatic extraction of data from corpora therefore proves to be potentially very fruitful for lexicography, since it adds a high number of WoCs that are not recorded in traditional dictionaries, even comprehensive ones such as DiCI. Human evaluation also confirms the complementarity of the two systems, since out of the total number of *valid* WoCs extracted by the two systems and not recorded in DiCI (534), only 80 combinations overlap.

These findings make us all the more convinced of the need for hybrid systems that simultaneously take into account information targeted in P-based and S-based approaches.

References

ATKINS, B.T.S. AND RUNDELL, M., 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

BARONI, M., BERNARDINI, S., COMASTRI, F., PICCIONI, L., VOLPI, A., ASTON, G. AND MAZZOLENI, M., 2004. Introducing the "La Repubblica" Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. *Proceedings of LREC 2004*, pp.1771–1774.

BENSON, M., BENSON, E. AND ILSON, R., 2010. *The BBI Combinatory Dictionary of English*. 3rd revised edition. Amsterdam/Philadelphia: John Benjamins.

CALZOLARI, N., FILLMORE, C.J., GRISHMAN, R., IDE, N., LENCI, A., MACLEOD, C. AND ZAMPOLLI, A., 2002. Towards best practice for multiword expressions in computational lexicons. *Proceedings of LREC 2002*, pp.1934–1940.

GRIES, S. TH., 2008. Phraseology and linguistic theory: a brief survey. In: S. Granger and F. Meunier, eds. *Phraseology: an interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins. pp.3–25.

HANKS, P., 2012. Corpus evidence and Electronic Lexicography. In: S. Granger and M. Paquot eds. *Electronic Lexicography*. Oxford: Oxford University Press. pp.57–82.

HEID, U. 2015. *Extracting linguistic knowledge about collocation from corpora*. Plenary talk delivered at the EUROPHRAS 2015 Conference. Malaga, June 29 – July 1, 2015.

LENCI, A., LAPESA, G. AND BONANSINGA, G., 2012. LexIt: A Computational Resource on Italian Argument Structure. *Proceedings of LREC 2012*, pp.3712–3718.

LENCI, A., LEBANI, G.E., CASTAGNOLI, S., MASINI, F. AND NISSIM, M., 2014. SYMPAThy: Towards a comprehensive approach to the extraction of Italian Word Combinations. *Proceedings of CLiC-it 2014* (Pisa, 9-11 December 2014), pp.234-238.

LENCI, A., LEBANI, G.E., SENALDI, M.S.G., CASTAGNOLI, S., MASINI, F. AND NISSIM, M., 2015. Mapping the Constructicon with SYMPAThy: Italian Word Combinations between fixedness and productivity. *Proceedings of the NetWordS Final Conference* (Pisa, March 30-April 1, 2015), pp.144-149.

LO CASCIO, V., 2013. *Dizionario combinatorio italiano*. Amsterdam/Philadelphia: John Benjamins.

MASINI, F., 2012. *Parole sintagmatiche in italiano*. Roma: Caissa.

NISSIM, M., CASTAGNOLI, S., MASINI, F. 2014. Extracting MWEs from Italian corpora: A case study for refining the POS-pattern methodology. *Proceedings of the 10th Workshop on Multiword Expressions* (MWE2014), pp.57–61.

PASSARO, L. C. AND LENCI, A. 2015. *Extracting Terms with EXTra*. Paper presented at the EUROPHRAS 2015 Conference. Malaga, June 29 – July 1, 2015.

PIUNNO, V., MASINI, F. AND CASTAGNOLI, S., 2013. *Studio comparativo dei dizionari combinatori dell'italiano e di altre lingue europee*. CombiNet Technical Report. Roma Tre University and University of Bologna.

RAMISCH, C. VILLAVICENCIO, A. AND BOITET, C., 2010. mwetoolkit: a framework for multiword expression identification. *Proceedings of LREC 2010*, pp.662–669.

SAG, I.A., BALDWIN, T., BOND, F., COPESTAKE, A. AND FLICKINGER, D., 2002. Multiword expressions: A pain in the neck for NLP. *Proceedings of CICLing 2002*, pp.1–15.

SQUILLANTE, L., 2015. *Polirematiche e collocazioni dell'italiano. Uno studio linguistico e computazionale*. Ph.D. dissertation Università di Roma "La Sapienza".

SERETAN, V., 2011. *Syntax-based collocation extraction*. Dordrecht: Springer.

VILLAVICENCIO, A., KORDONI, V., ZHANG, Y., IDIART, M. AND RAMISCH, C., 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp.1034–1043.

VOGHERA, M., 2004. Polirematiche. *Linguistica Pragensia*, 67(2), pp.100–108.