Automatic extraction of Word Combinations from corpora: evaluating methods and benchmarks

Malvina Nissim[1], Sara Castagnoli[2], Francesca Masini[2], Gianluca E. Lebani[3], Lucia Passaro[3], Alessandro Lenci[3]
[1]CLCG, University of Groningen, [2]University of Bologna, [3]University of Pisa
m.nissim@rug.nl, {s.castagnoli,francesca.masini}@unibo.it,
{gianluca.lebani,lucia.passaro}@for.unipi.it,
alessandro.lenci@unipi.it

Abstract

English. We report on three experiments aimed at comparing two popular methods for the automatic extraction of Word Combinations from corpora, with a view to evaluate: i) their efficacy in acquiring data to be included in a combinatory resource for Italian; ii) the impact of different types of benchmarks on the evaluation itself.

Italiano. *Presentiamo i risultati di tre esperimenti che mirano a confrontare due metodi di estrazione automatica di combinazioni di parole da corpora, con lo scopo di: (i) valutare l'efficacia dei due metodi per acquisire dati da includere in una risorsa combinatoria per l'italiano, e (ii) analizzare e confrontare i metodi di valutazione stessi.*

1 Introduction

We use the term **Word Combinations** (WoCs) to encompass both Multiword Expressions, namely WoCs characterised by different degrees of fixedness and idiomaticity, such as idioms, phrasal lexemes, collocations, preferred combinations (Calzolari et al., 2002; Sag et al., 2002; Gries, 2008), and the distributional properties of a word at a more abstract level (argument structure, subcategorization frames, selectional preferences).

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 (AMs) 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**) (Lenci et al., 2014; Lenci et al., 2015).

While P-based approaches have shown to yield satisfactory results for relatively fixed, short and adjacent WoCs, it has been suggested that syntactic dependencies might be more helpful to capture discontinuous and syntactically flexible WoCs (Seretan, 2011). The two methods intuitively seem to be highly complementary rather than competing with one another, and attempts are currently being proposed to put them together (Lenci et al., 2014; Lenci et al., 2015; Squillante, 2015). In previous work (Castagnoli et al., forthcoming), we compared the performance of the two methods against two benchmarks (a dictionary and expert judgments), showing that the two methods are indeed complementary and that automatic extraction from corpora adds a high number of WoCs that are not recorded in manually compiled dictionaries.

As an extension of that work, in this paper we shift the focus of investigation by addressing the following research questions: What is the effect of different benchmarks when evaluating an extraction method? What do our results tell us about the bechmarks themselves? And, as a byproduct, can experts / laypeople be exploited to populate a lexicographic combinatory resource for Italian?

2 Benchmarks

The performance of WoC extraction can be evaluated in various ways. A straightforward way is assessing extracted combinations against an existing dictionary of WoCs (Evaluation 1). Such resources, however, are often compiled manually on the basis of the lexicographers' intuition only. The dictionary can be seen as a one-expert judgement, in a top-down (lexicographic) fashion. Moreover, this type of evaluation assumes the dictionary as an absolute gold standard, without considering that any dictionary is just a partial representation of the

lexicon and that corpus-based extraction might be able to identify further possible WoCs.

Another way to assess the validity of extracted combinations is via human evaluation. One problem with this approach lies in the competence of the judges: experts are difficult to recruit, but it isn't completely clear whether people unfamiliar with linguistic notions are able to grasp the concept of WoCs, and to judge the validity of the extracted strings. Knowing whether this is a task that can be performed by laypeople is not only theoretically interesting, but also practically useful. To this end, we set up two distinct human-based experiments: one involving experts (Evaluation 2), and one involving laypeople (Evaluation 3). Table 1 summarises the characteristics of the three strategies, whose results are discussed and compared in the next sections, in terms of the kind and number of contributors, the procedure (bottom-up means that the evaluation is done directly on the corpus-extracted WoCs rather than against a pre-compiled list (top-down)), the assessment performed or required, and the data evaluated.

3 Experimental evaluation

3.1 Data and WoC extraction

We selected a sample of 25 Italian target lemmas (TLs) – 10 nouns, 10 verbs and 5 adjectives – and we extracted P-based and S-based combinatory information from *la Repubblica* corpus (Baroni et al., 2004)[1]. TLs were selected by combining frequency information derived from *la Repubblica* and inclusion in DiCI (Lo Cascio, 2013), a manually compiled dictionary of Italian WoCs, which is also used for (part of the) evaluation.

As regards the P-based method, we extracted all occurrences of each TL in a set of 122 predefined POS-patterns deemed representative of Italian WoCs, using the **EXTra** tool (Passaro and Lenci, forthcoming). EXTra retrieves all occurrences of the specified patterns as linear and contiguous sequences (no optional slots) and ranks them according to various association measures, among which we chose Log Likelihood (LL). The search considers lemmas, not wordforms. Only sequences with frequency over 5 were considered.

As regards the S-based method, we extracted the distributional profile of each TL using the **LexIt** tool (Lenci et al., 2012). 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. 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. Only candidate WoCs with frequency over 5 have been considered.

3.2 Evaluation against a dictionary

The gold standard we used for this part of the evaluation, fully presented in (Castagnoli et al., forthcoming), is the DiCI dictionary (Lo Cascio, 2013).

Recall is calculated as the percentage of extracted candidates out of the combinations found in the gold standard. Generally, EXTra performs better than LexIt for nominal and adjectival TLs, whereas LexIt has a higher recall for virtually all verbal TLs.[2] **R-precision**, which measures precision at the rank position corresponding to the number of combinations found in DiCI, is almost always higher for LexIt than for Extra, irrespective of POS. Total **overlap** is calculated as the percentage of cases in which EXTra/LexIt retrieve (or not) the same gold standard combinations. For instance, the entry for *giovane* 'young' in DiCI contains 50 combinations. Out of these, 20 are retrieved by both EXTra and LexIt, 27 are retrieved by neither, and only LexIt extracts 3 further WoCs. This means that the two systems perform similarly for 94% of cases found in the benchmark data. Total overlap runs between 59.07% and 94% (average 76.05%).

3.3 Human-based evaluation with experts

We recruited a number of linguists, mainly with a background in translation and/or corpus work. They were asked to assess the validity of candidates by assigning one of 3 possible values: Y (Yes, a valid WoC), N (No, not a valid WoC), U (Uncertain / may be part of a valid WoC). We obtained judgments for 2,000 candidates (50% EXTra, 50% LexIt, taking the top 100 results for 10 TLs from each system). We used two annotators per

[1]The version we used was POS-tagged with the tool described in (Dell'Orletta, 2009) and dependency-parsed with DeSR (Attardi and Dell'Orletta, 2009).

[2]This result may in part be due to the POS-patterns used, which were limited to a maximum of 4 slots, thus preventing EXTra from capturing longer verbal expressions. However, this can be seen as an inherent limitation of the P-based approach, given that the complexity/variability of patterns increases immensely as soon as we consider longer strings.

Table 1: Overview of evaluation strategies.

	Evaluation 1 (DiCI)	Evaluation 2 (experts)	Evaluation 3 (laypeople)
contributors	expert (1)	expert (> 1)	naive (> 1)
procedure	top-down	bottom-up	bottom-up
assessment	inclusion	validity (categorical)	typicality + idiomaticity (scalar)
candidates	all extracted (ca.105,000)	top extracted per TL (2,000)	random from Eval 2 (630)

candidate, and considered valid WoCs those that received either YY or YU values.

A total of 855 entries (EXTra: 408, LexIt: 447) were judged as valid. Out of these, 534 (62.5%) are not recorded in DiCI (EXTra: 273, LexIt: 261). If we intersect the two sets, we find that only 80 of these additional WoCs are in common, which means that we have 454 actual *new* valid WoCs, retrieved thanks to the corpus-based methodology.

3.4 Human-based evaluation with laypeople

Judgements from laypeople were obtained by setting up a crowdsourcing task on the Crowdflower platform (`http://www.crowdflower.com`). Compared to the previous experiment, annotators were asked to judge two aspects of the candidate combinations: how *typical* they are, i.e. how important it is that they are included in a multiword dictionary; and how *idiomatic* they are, i.e. how much their overall meaning is not directly inferrable from their parts (non-compositionality). Both judgements were asked on a scale from 1 to 5 rather than via the discrete values used by the experts (Y/N/U). The Appendix shows a snapshot of the instructions and the task the annotators were presented with. Note that candidates were presented in the form they were extracted from the corpora, i.e. lemmatized (e.g. *vero guerra* instead of *vera guerra* 'true war'). Further, LexIt examples may contain free slots (e.g. *pagare * multa* 'pay * fine').

This second human-based experiment was primarily expected to shed light on whether experts' and laypeople's judgements differ in the assessment of WoCs. Moreover, the additional question about idiomaticity was aimed at detecting potential differences in the degree of idiomaticity of the WoCs the two methods extract.

3.4.1 Participation and results

Potential annotators could train on some "gold" combinations, which were also used to assess the quality of the contributors. Such gold combinations were not part of the original dataset and are not further included in the analysis. Contributors who misclassified more than 60% of the test questions were not allowed to proceed with the rest of the combinations, so that out of 81 potential contributors we were left with 53 reliable ones, and only 36 actively working on the task (with contributions ranging from 300 to 20 annotated combinations).

As a result, this second human-based experiment is based on 630 combinations (a random subsample of the original 2,000 dataset of the expert-based evaluation) for which we managed to collect three independent judgements. The distribution between combinations extracted by Extra (322) and by Lexit (308) is approximately preserved.

In Figure 1 and 2 we report the results of the evaluation for the "typicality" and "idiomaticity" assessments (x in the chart labels), respectively, splitting the overall range into five subranges.

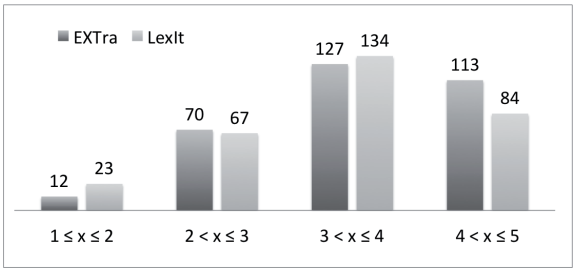


Figure 1: Results of the crowdsourcing evaluation for how *typical* combinations are (average of three annotations, global range 1–5).

If we deem *valid* any combination with average score > 3 (the two rightmost columns in the Figures), we can observe that laypeople judged as valid combinations the majority of candidates in both sets and more precisely: approx. 75% of candidates extracted by EXTra (240/322) and approx. 71% of candidates extracted by LexIt (218/308). The two methods perform similarly also regarding the capability of extracting combinations with stronger or weaker idiomaticity: approx. 38% of (those judged as) typical combinations obtained via

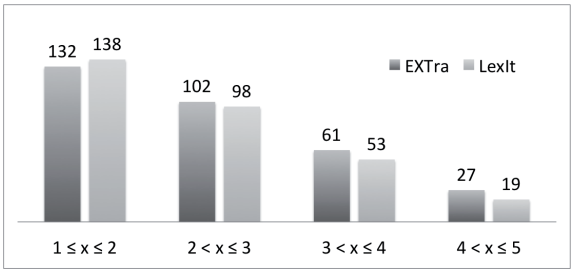


Figure 2: Results of the crowdsourcing evaluation for how *idiomatic* combinations are (average of three annotations, global range 1–5).

Table 2: Comparison of "valid" combinations according to laypeople and expert judges.

	valid for both	laypeople only	total
EXTra	124	116	240
LexIt	119	99	218

EXTra were also judged idiomatic (88), and approx. 33% of (those judged as) typical combinations obtained via LexIt were also judged idiomatic (72). Overall, EXTra appears to have a slightly better performance in both cases (although the difference is not statistically significant), and this is different from what we observed in the expert-based evaluation. The reason for this may lie in the fact that the LexIt candidates correspond to more abstract and schematic WoCs, which could eventually be harder to map onto specific instances by the evaluators.

3.4.2 Experts vs laypeople

Do experts and laypeople share the same notion of what a *typical* combination is? Given that in the crowdsourcing experiment we used a subset of the expert set, we checked how many of those combinations that were assessed as valid (> 3) by laypeople had also been evaluated as valid by the experts (YY or YU, see Section 3.3). Table 2 shows the results of such comparison. If we treat the experts' judgements as gold, we can interpret the values in the table as precision, resulting in 0.517 for EXTra and 0.546 for LexIt. Both figures are rather low, and suggest that the notion of "typicality" of a combination - or possibly the notion of a combination at all - isn't at all straightforward.

A qualitative analysis of the disagreements between laypeople and experts leads to some interesting insights. Combinations annotated as valid only by the former include: a) cases were the candidate differs from a proper WoC only for a small detail: e.g. *dichiarare una guerra* 'declare a war' (proper WoC: *dichiarare guerra* 'declare war', without indefinite article), *tenere il ostaggio* 'take the hostage' (proper WoC: *tenere in ostaggio* 'take s.one hostage'), showing little attention to details; b) cases of uncertain collocations: e.g. *libretto rosso* 'red booklet', *famiglia italiano* 'Italian family', *prendere - carta* 'take - paper'; c) blatantly incomplete/nonsensical combinations: e.g. *di guerra di* 'of war of', *di molto famiglia* 'of many family'; d) a few WoCs that were not recognised as valid by experts: e.g. *dare la mano* 'shake one's hand', *prendere corpo* 'to take shape', *guerra punica* 'punic war'.

4 Discussion and conclusion

As for extraction methods per se, we observed that recall against a manually compiled WoC dictionary is good for both EXTra and LexIt, and, especially, that the two systems are complementary. In the human evaluation performed by experts, 40% of WoCs automatically extracted with EXTra and LexIt are deemed valid, and more than half of these are not attested in DiCi. We can thus say that data from corpora proves to be very fruitful, especially if we use the two methods complementarily.

As for benchmarks, we observed that the dictionary we have evaluated is not an exhaustive resource, and should be complemented with corpus-extracted WoCs. We also observed that expert- and laypeople-based evaluations differ, which raises a number of interesting, albeit puzzling questions. Overall, it seems that the notion of WoC, as well as of idiomaticity, is quite a complex one to grasp for non-linguists: the collection of judgments took quite a long time to be completed (much more than we expected) and evaluators explicitly regarded the task and the instructions as particularly complex.

The results of our experiments thus leave us with a sort of methodological conundrum, as both a dictionary-based gold standard and a human-based evaluation have limitations. Using experts not only makes the evaluation expensive, but also little ecological, as it is standard practice in psycholinguistics and computational linguistics to resort to laypeople judgments. The fact that evaluating WoCs isn't easy for laypeole may cast some shadows on the concept of WoC itself. This suggests that improving extraction methods must go hand in hand with the theoretical effort of making the very notion of WoC more precise, in order to make it an experimentally solid and testable notion.

Acknowledgments

This research was carried out within the **CombiNet** project (PRIN 2010-2011 *Word Combinations in Italian: theoretical and descriptive analysis, computational models, lexicographic layout and creation of a dictionary*, n. 20105B3HE8), funded by the Italian Ministry of Education, University and Research (MIUR). `http://combinet.humnet.unipi.it`.

References

Giuseppe Attardi and Felice Dell'Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In *Proceedings of NAACL 2009*, pages 261–264.

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 *Proceedings of LREC 2004*, pages 1771–1774.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of LREC 2002*, pages 1934–1940.

Sara Castagnoli, Gianluca E. Lebani, Alessandro Lenci, Francesca Masini, Malvina Nissim, and Lucia C. Passaro. forthcoming. Pos-patterns or syntax? comparing methods for extracting word combinations. In *Proceedings of EUROPHRAS 2015 (provisional)*.

Felice Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. In *Proceedings of EVALITA 2009*.

Stefan Th. Gries. 2008. Phraseology and linguistic theory: a brief survey. In Sylviane Granger and Fanny Meunier, editors, *Phraseology: an interdisciplinary perspective*, pages 3–25. John Benjamins, Amsterdam & Philadelphia.

Alessandro Lenci, Gabriella Lapesa, and Giulia Bonansinga. 2012. LexIt: A Computational Resource on Italian Argument Structure. In *Proceedings of LREC 2012*, pages 3712–3718.

Alessandro Lenci, E. Gianluca Lebani, Sara Castagnoli, Francesca Masini, and Malvina Nissim. 2014. SYMPAThy: Towards a comprehensive approach to the extraction of Italian Word Combinations. In *Proceedings of CLiC-it 2014*, pages 234–238, Pisa, Italy.

Alessandro Lenci, E. Gianluca Lebani, S.G. Marco Senaldi, Sara Castagnoli, Francesca Masini, and Malvina Nissim. 2015. Mapping the Constructicon with SYMPAThy: Italian Word Combinations between fixedness and productivity. In *Proceedings of the NetWordS Final Conference*, pages 144–149, Pisa, Italy.

Vincenzo Lo Cascio. 2013. *Dizionario Combinatorio Italiano (DiCI)*. John Benjamins, Amsterdam/Philadelphia.

Lucia C. Passaro and Alessandro Lenci. forthcoming. Extracting terms with EXTra. In *Proceedings of EUROPHRAS 2015 (provisional)*.

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.

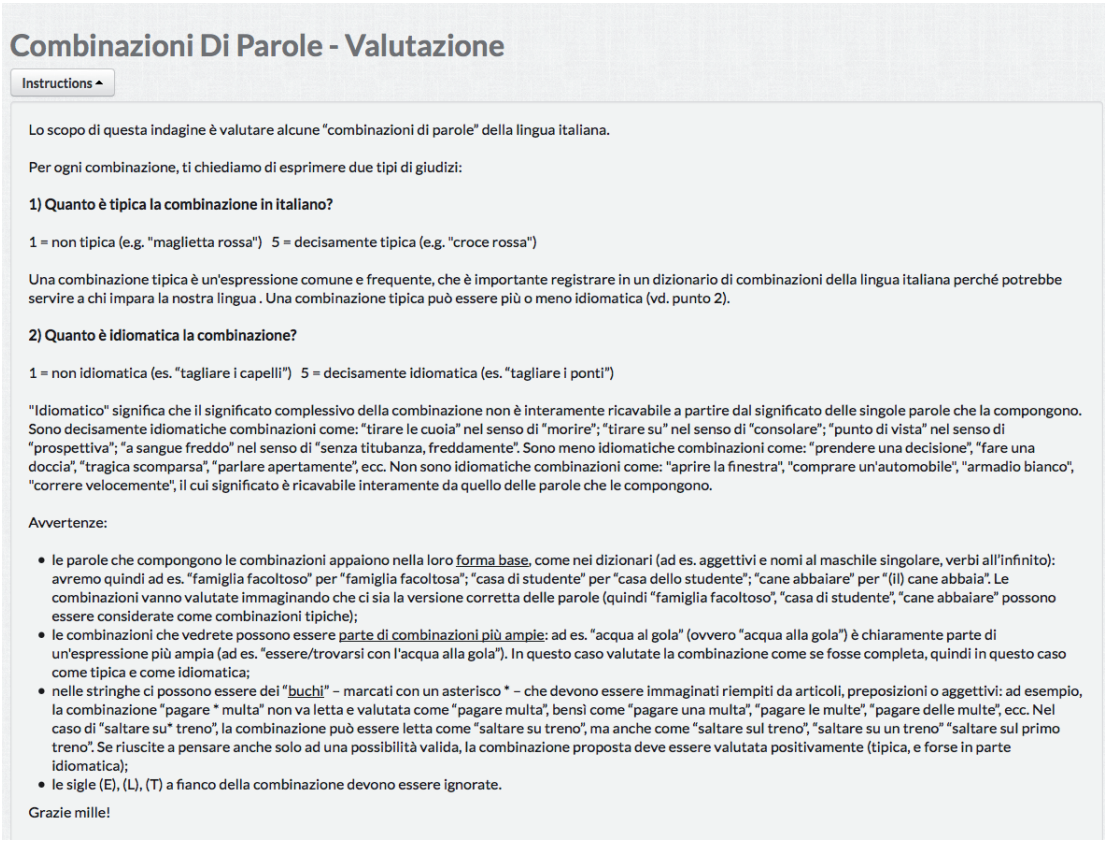
Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing 2002*, pages 1–15.

Violeta Seretan. 2011. *Syntax-based collocation extraction*. Springer, Dordrecht.

Luigi Squillante. 2015. *Polirematiche e collocazioni dell'italiano. Uno studio linguistico e computazionale.* Ph.D. thesis, Università di Roma "La Sapienza".

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of EMNLP-CoNLL 2007*, pages 1034–1043.

Appendix: Crowdflower Job



Combinazioni Di Parole - Valutazione

Instructions

Lo scopo di questa indagine è valutare alcune "combinazioni di parole" della lingua italiana.

Per ogni combinazione, ti chiediamo di esprimere due tipi di giudizi:

1) Quanto è tipica la combinazione in italiano?

1 = non tipica (e.g. "maglietta rossa") 5 = decisamente tipica (e.g. "croce rossa")

Una combinazione tipica è un'espressione comune e frequente, che è importante registrare in un dizionario di combinazioni della lingua italiana perché potrebbe servire a chi impara la nostra lingua . Una combinazione tipica può essere più o meno idiomatica (vd. punto 2).

2) Quanto è idiomatica la combinazione?

1 = non idiomatica (es. "tagliare i capelli") 5 = decisamente idiomatica (es. "tagliare i ponti")

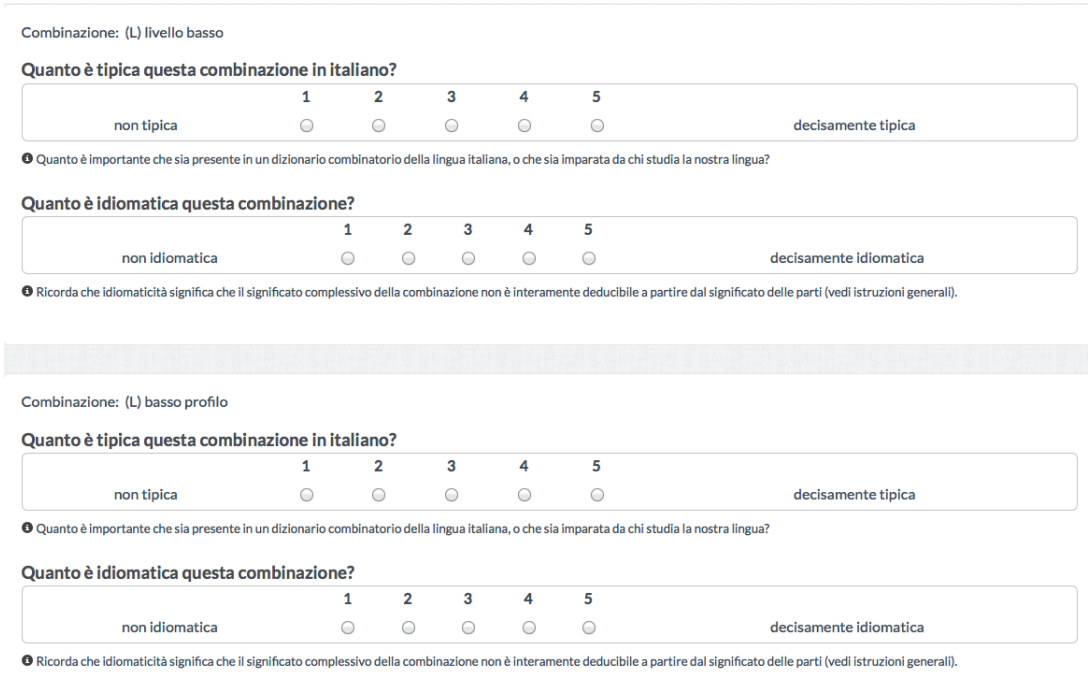
"Idiomatico" significa che il significato complessivo della combinazione non è interamente ricavabile a partire dal significato delle singole parole che la compongono. Sono decisamente idiomatiche combinazioni come: "tirare le cuoia" nel senso di "morire"; "tirare su" nel senso di "consolare"; "punto di vista" nel senso di "prospettiva"; "a sangue freddo" nel senso di "senza titubanza, freddamente". Sono meno idiomatiche combinazioni come: "prendere una decisione", "fare una doccia", "tragica scomparsa", "parlare apertamente", ecc. Non sono idiomatiche combinazioni come: "aprire la finestra", "comprare un'automobile", "armadio bianco", "correre velocemente", il cui significato è ricavabile interamente da quello delle parole che le compongono.

Avvertenze:

- le parole che compongono le combinazioni appaiono nella loro forma base, come nei dizionari (ad es. aggettivi e nomi al maschile singolare, verbi all'infinito): avremo quindi ad es. "famiglia facoltoso" per "famiglia facoltosa"; "casa di studente" per "casa dello studente"; "cane abbaiare" per "(il) cane abbaia". Le combinazioni vanno valutate immaginando che ci sia la versione corretta delle parole (quindi "famiglia facoltoso", "casa di studente", "cane abbaiare" possono essere considerate come combinazioni tipiche);
- le combinazioni che vedrete possono essere parte di combinazioni più ampie: ad es. "acqua al gola" (ovvero "acqua alla gola") è chiaramente parte di un'espressione più ampia (ad es. "essere/trovarsi con l'acqua alla gola"). In questo caso valutate la combinazione come se fosse completa, quindi in questo caso come tipica e come idiomatica;
- nelle stringhe ci possono essere dei "buchi" – marcati con un asterisco * – che devono essere immaginati riempiti da articoli, preposizioni o aggettivi: ad esempio, la combinazione "pagare * multa" non va letta e valutata come "pagare multa", bensì come "pagare una multa", "pagare le multe", "pagare delle multe", ecc. Nel caso di "saltare su* treno", la combinazione può essere letta come "saltare su treno", ma anche come "saltare sul treno", "saltare su un treno" "saltare sul primo treno". Se riuscite a pensare anche solo ad una possibilità valida, la combinazione proposta deve essere valutata positivamente (tipica, e forse in parte idiomatica);
- le sigle (E), (L), (T) a fianco della combinazione devono essere ignorate.

Grazie mille!

Screenshot of the Crowdflower job: instructions.



Combinazione: (L) livello basso

Quanto è tipica questa combinazione in italiano?

	1	2	3	4	5	
non tipica	○	○	○	○	○	decisamente tipica

Quanto è importante che sia presente in un dizionario combinatorio della lingua italiana, o che sia imparata da chi studia la nostra lingua?

Quanto è idiomatica questa combinazione?

	1	2	3	4	5	
non idiomatica	○	○	○	○	○	decisamente idiomatica

Ricorda che idiomaticità significa che il significato complessivo della combinazione non è interamente deducibile a partire dal significato delle parti (vedi istruzioni generali).

Combinazione: (L) basso profilo

Quanto è tipica questa combinazione in italiano?

	1	2	3	4	5	
non tipica	○	○	○	○	○	decisamente tipica

Quanto è importante che sia presente in un dizionario combinatorio della lingua italiana, o che sia imparata da chi studia la nostra lingua?

Quanto è idiomatica questa combinazione?

	1	2	3	4	5	
non idiomatica	○	○	○	○	○	decisamente idiomatica

Ricorda che idiomaticità significa che il significato complessivo della combinazione non è interamente deducibile a partire dal significato delle parti (vedi istruzioni generali).

Screenshot of the Crowdflower job: examples involving the TL *basso* 'low/short'.