Combining Grammatical and Relational Approaches. A Hybrid Method for the Identification of Candidate Collocations from Corpora

Damiano Perri[*], **Irene Fioravanti**[†], **Osvaldo Gervasi**[*], **Stefania Spina**[†]
[*]University of Perugia
[†]University for Foreigners of Perugia
damiano.perri, osvaldo.gervasi@unipg.it
irene.fioravanti, stefania.spina@unistrapg.it

Abstract

We present an evaluation of three different methods for the automatic identification of candidate collocations in corpora, part of a research project focused on the development of a learner dictionary of Italian collocations. We compare the commonly used POS-based method and the syntactic dependency-based method with a hybrid method integrating both approaches. We conduct a statistical analysis on a sample corpus of written and spoken texts of different registers. Results show that the hybrid method can correctly detect more candidate collocations against a human annotated benchmark. The scores are particularly high in adjectival modifier relations. A hybrid approach to candidate collocation identification seems to lead to an improvement in the quality of results.

Keywords: Collocation, Automatic identification, Learner dictionary

1. Introduction

Multi-word expressions (henceforth, MWEs), defined as lexical units (collocations, idioms, lexical bundles, etc.) consisting of two or more words, have been the focus of extensive research in many areas including lexicography and NLP for several decades (Evert, 2004; Paquot, 2015; Spina, 2020). The creation of lexicographical combinatory resources, such as dictionaries of collocations, explicitly targeted to learners of second languages (L2s), has been undertaken mainly for English (McIntosh et al., 2002); (Rundell, 2010), although general dictionaries of collocations not explicitly addressed to L2 learners exist for several languages, including English (Benson et al., 1986), and Italian (Urzì, 2009; Tiberii, 2012; Lo Cascio, 2013). The use of language corpora has significantly boosted research on MWEs and their lexicographic applications. This is particularly evident in the area of lexicography dedicated to MWEs, where the identification of typical word combinations hugely benefits from the use of vast collections of texts. These corpora allow to extract frequent naturally occurring lexical patterns, with the aid of NLP and statistical techniques for the analysis of word combinations (Hanks, 2012).

Two main tasks are involved in the extraction of MWEs from corpora (Seretan, 2011): the automatic identification of candidates, often according to specific a priori criteria on their grammatical and/or syntactic patterns, and the detection of phraseologically meaningful combinations (collocations, in this case), often based on frequency and/or statistical association measures, to filter out sequences of words without phraseological relevance. In this study, our focus is on the first task of automatically identifying candidate collocations in Italian corpora. We assume that the effectiveness of the subsequent stages in creating a learner dictionary of collocation strongly depends on how accurate this candidate identification proves to be. The more an automatic system based on NLP techniques can accurately identify word combinations that are potential collocations, the more accurate the data on their frequency. As a consequence, the association measures used to filter out non-collocations, all of which are, to varying degrees, dependent on frequency, can benefit from more reliable frequency values, resulting in increased accuracy.

The present study reports on an experiment aimed at proposing a hybrid approach to this task by comparing and evaluating the two most commonly used candidate detection methods - the POS-based method and the syntactic dependency-based method - with a third one resulting from the integration of the two previous approaches. For the first two methods, we adopt the denomination from (Castagnoli et al., 2016) and refer to the POS-based as the P-based approach and the dependency-based as the S-based approach, while we refer to the third integrated method as the Hybrid approach. Current collocation extraction approaches rely on linguistic pre-processing (e.g., POS-tagging or dependency parsing) of source corpora to better identify the candidates (Seretan, 2011). Previous research has shown that the P-based and S-based approaches have some limitations. The former re-

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@*LREC-COLING-2024*, pages 138–146. May 25, 2024.

lies on an accurate and established NLP task such as POS-tagging. However, relying on positional POS patterns, it fails to capture the syntactic relations between word pairs or the marked sentence structures where the regular constituent order is reversed. For instance, a P-based approach would not detect the verb-direct object relation between *play* and *role* in Example 1 (the example is taken from Seretan, 2011, 59).

Example 1. *It is true, we must combat the menace of alcoholism in young people, and this text successfully highlights the* ***role*** *that families, teachers, producers and retailers must* ***play*** *in this area.*

On the contrary, this relation would probably be detected using an S-based approach that relies on parsed data and thus can identify the verb-direct object dependency. Another advantage of this approach is that it does not limit the distance between the two words constituting the candidate collocation, unlike the P-approach. However, parsing errors are a well-known shortcoming of this approach: error rates ranging from 7.85% to 9.7% of the total candidate collocations extracted were reported to be due to parsing errors by previous studies (Wu and Zhou, 2003; Lin, 1999). Despite the recent improvement in parsing accuracy, (Qi et al., 2020; Akbik et al., 2018) the parsing approach still has limitations in selecting candidate collocations as it provides little information on how words combine with each other and fail to distinguish frequent combinations and idiomatic ones with the same syntactic structure (Castagnoli et al., 2016).

This study aims to present a hybrid approach to detecting candidate collocation from corpora for lexicographic applications on a language different from English, i.e. Italian. The hypothesis we aim to validate is that this hybrid approach performs better in the candidate identification task. From an exploratory perspective, we also intend to investigate cases in which the hybrid method works better and identify cases where further improvements might be warranted.

2. Related work

In this section, we briefly survey the main methods and NLP techniques used to perform the specific task of detecting, or discovering (Constant et al., 2017) candidate collocations from corpora, regardless of the measures employed to identify the proper phraseological collocations, which represents a further step in the process of assembling the set of entries required by the lexicographic application.

Early NLP works addressing this task identified candidate collocations using frequent word sequences, regardless of their syntactic structure, and relied on n-gram methods to extract them from corpora (Choueka, 1988; Smadja, 1993). Later, this search "for needles in a haystack" (Choueka, 1988) more and more employed linguistically preprocessed corpora and lemmatised and POS-tagged data. This further step was especially suitable for handling morphological and syntactic variability typical of languages with richer morphology and more accessible word order (Evert, 2004). The P-approach is the first to become established, given the widespread availability of POS-tagged corpora in many languages. Many extraction systems relying on this approach involve an a priori selection of specific types of POS combinations (e.g. verb-noun, adjective-noun, etc.). Right from the start, a drastic improvement in the detection accuracy was found when a POS filter was applied (Breidt, 1993; Daille, 1994; Krenn, 2000; Ritz, 2006). These results were primarily reported for fixed and adjacent candidates, where even a simple linguistic analysis can capture basic grammatical patterns.

In later years, it has been suggested that the detection of candidate collocations can benefit from a finer linguistic analysis of texts. Seretan's (2011) extensive study explored and evaluated the use of syntactic dependencies, as they can also capture discontinuous and syntactically flexible candidate collocations based on syntactic relations between words, improving the quality of the results. However, many systems relying on an S-approach aimed at MWE identification after parsing, so as to benefit from the previous syntactic analysis (Constant et al., 2017) reported high parsing error rates affecting the accuracy of the detection task. The issue of parsing accuracy is identified and evaluated by several studies (e.g. Orliac and Dillinger, 2003; Lü and Zhou, 2004). Lü and Zhou (2004) identified a parsing error rate >7%. Orliac and Dillinger (2003) also evaluated the most recurrent parsing errors and found that relative constructions were responsible for nearly half of the candidate collocations missed by their system.

Given all these reported limitations, it can be argued that the existing detection methods relying on an S-based approach are promising but have not yet been fully developed, due to issues related to parsing accuracy. There is, therefore, a general call for hybrid approaches to candidate collocation detection, combining the advantages of both P-based and S-based approaches while minimising their shortcomings. As Castagnoli et al. (2016) claimed, "the two methods seem to be highly complementary rather than competing with one another". Some attempts have been made to integrate the two approaches in recent years. Simkó et al. (2017) proposed a system using both POS-tagging and dependency parsing to identify single- and multi-token verbal MWEs in texts and reported the best results on the verb-particle constructions where their sys-

tem correctly identified around 60% of constructions, but only about 40% of other types. Shi and Lee (2020) proposed a joint method that combines scores from both POS-tagging and dependency parsing to extract headless MWEs. Their results showed that tagging is more accurate than parsing for identifying flat-structure MWEs. At the same time, the joint method leads to higher accuracy, and most of the gains derive from shared results between parsers and taggers.

3. Method

To validate our hypothesis and explore the performance of different systems in automatically detecting candidate collocations in Italian corpora, we designed our experiment to mimic the "natural" processes that will be employed in the final extraction of candidates to be included in a learner dictionary of Italian collocations. For instance, we did not preselect target words or lemmas for the experiment. Instead, we considered all the word pairs produced in a text sample.

The only pre-selection we made was the syntactic relations of the candidate collocations. We opted to focus on syntactically-bound combinations, as the task of detecting candidate collocations is targeted to a lexicographic application. In the final dictionary entries, these collocations will be presented in accordance with their syntactic patterns. The choice was to investigate the two dependencies verb + direct object (Vdobj) and adjective modifier (amod) before and after a noun (both word orders are allowed in Italian). The choice is motivated by reasons of coverage and diversification. Firstly, previous research has shown that, among the eight syntactic structures most commonly forming collocations in Italian (verb + direct object, amod, noun + preposition + noun, noun + noun, verb + adjective, verb + adverb, noun + conjuction + noun, adjective + conjunction + adjective), the two that are considered in this study (Vdobj and amod) cover more than 50% of the total structures (Spina, 2016). Furthermore, while in both relations the order of the two components can be reversed, they have different features in terms of distance between their two components. In the Vdobj word combinations the distance between the two components can be even of several words (Example 2: there are five words between the verb *mantenere* 'keep' and the direct object *promesse* 'promises', and the two words are connected by a relative pronoun), while in the case of amod the two words are usually adjacent (Example 3) or near adjacent (Example 4).

Example 2. Non fare **promesse** che non riuscirai mai a **mantenere**!
Don't make promises you will never keep!

Example 3. Elisa mi stava raccontando della sua **brutta avventura**
Elisa was telling me about her bad adventure

Example 4. Questo è il **momento** più **atteso** della giornata
This is the most awaited moment of the day

3.1. Sample texts

We randomly extracted eight texts from a reference corpus of Italian, the *Perugia corpus* (Spina, 2014; https://lt.eurac.edu/cqpweb/), of the total size of ca. 8,000 tokens, balanced across written (tokens = 4,000) and spoken (tokens = 4,000) registers. We included different text genres: two newspaper articles (a report and an editorial), two school essays and a tourism-related blog post for the written part, and transcriptions of a conference, of a political speech and of the dialogues of a television series for the spoken part. On the one hand, this diversification in registers and text genres allows us to perform a simulation close to the actual extraction of candidate collocations for all the combination types in the whole corpus. On the other hand, it enables us to evaluate the three approaches to this task for register variation, which could affect accuracy.

3.2. The three systems

We used the systems described below to compare three different methods for detecting and extracting candidate collocations from Italian corpora, whose output was compared with a benchmark of human annotation.

P-based approach The sample texts were POS-tagged using *TreeTagger* (Schmid, 1994), trained with an ad hoc tagset based on a fine-grained set of 54 POS tags (Spina, 2014). Afterwards, the texts were searched via the *Corpus Workbench* (CWB) tool (Hardie, 2012) and the *Corpus Query Processing* (CQP) system by using three separate queries to detect the Vdobj relations and the two positional variants of the amod relations, with the adjective preceding or following the modified nouns. The three queries integrate POS tag sequences (the target ADJ, NOUN and VERB POS tags, as well as those that can potentially be inserted within the two constituents of the combinations, like articles, conjunctions or adverbs) and regex with lemmas to exclude (a list of the most frequent intransitive Italian verbs). The direct output of this regex-over-pos process represents the P-based approach, that was able to identify 549 candidate collocations.

S-based approach In this approach, a candidate collocation consist of two syntatically related lexical

items. Therefore, the main criterion for detecting a candidate is the presence of a syntactic relation between the two items, in our case, the Vdobj and amod relations. In addition, to be identified as a valid candidate, each pair must satisfy more specific grammatical constraints. For instance, the words involved in the syntactic relations can only be nouns, adjectives or verbs. The sample texts were parsed using the framework of Universal Dependencies for treebank annotation (UD; de Marneffe et al., 2021) and the popular open-source library for advanced NLP in Python *spaCy*. Artificial intelligence is to date applied in many areas of science (Benedetti et al., 2020; Perri et al., 2022; Milani et al., 2021). The *spaCy* library is an example of the application of artificial intelligence to linguistic analysis. Since the simple parsing output does not yet represent the S-approach, the complete procedure details are described in section 3.3. The final number of candidate collocations identified by the S-based approach is 685.

Hybrid approach The hybrid approach results from merging the two previous approaches. It includes all the common candidates identified by both, as well as those only detected by the P-based approach and those only detected by the S-based approach. The Hybrid approach identified 748 candidate collocations.

3.3. Annotation

The output of the three systems was compared to a benchmark obtained by human evaluation. Two Italian trained linguists manually extracted all the Vdobj and amod combinations used in the eight sample texts. The two human annotators only adopted the criterion of the syntactic relations to extract the candidate collocations. Without calculating the inter-annotator agreement, any inter-annotator disagreements were resolved through negotiation until consensus was achieved for all forms. This annotation process resulted in a list of 610 candidate collocations, which served as a benchmark for the following steps.

3.4. Computational procedure

Three steps make up the computational process, allowing consistent and thorough data processing. The preliminary pre-processing of the texts was first carried out to enable homogeneous treatment of information. In the second step, the sentences were parsed using *spaCy*, and a set of rules was implemented to optimise the analysis. Finally, the results were statistically treated. Specifically, the results obtained through the S-approach were compared to those obtained through the P-approach and the Hybrid approach.

3.4.1. The pre-processing of the input texts

The first step involved pre-processing the texts to standardise the input data format and remove any irrelevant elements for analysis. This process included inserting capital letters at the beginning of each sentence and full stops at the end. We removed all whitespace due to typing errors (e.g. double whitespace) or whitespace after the end of a sentence in order to ensure that all input is as clean and error-free as possible. The sentences were then extracted and inserted into a data structure. Each sentence was assigned to a row within a spreadsheet (CSV file), constituting the database for the following stages of the analysis. Having one sentence per line is crucial, as it ensures an easily repeatable analysis and prevents overloading the *spaCy* parser, which can operate with a limited amount of RAM without requiring excessive resources.

3.4.2. The parsing of input phrases

The second phase of our work was devoted to sentence parsing using *spaCy* and the rules implemented in Python to recognize adjective modifier dependency (amod) and verb-direct object dependency (Vdobj).

The syntactic analyzer is a Python object obtained by importing the pre-trained *spaCy* library on the CPU-optimized Italian pipeline called `it_core_news_lg`[1]. The pre-training model occupies 541MB of written text (news and media). The pipeline provided by the model consists of `tok2vec`, `morphologizer`, `tagger`, `parser`, `lemmatizer`, `attribute_ruler`, `ner`. *spaCy* was trained with the UD Italian ISDT v2.8 (Italian Stanford Dependency Treebank; Attardi et al., 2015) There are various software libraries that can be used to perform the task of analysing the grammar of a sentence. We opted for *spaCy* since a version of its Italian language model was released very recently, on 1 Oct 2023[2].

Each sentence in our corpus was analyzed word by word. Given a word, *spaCy* provides a list of output objects: DepRel, `Form`, `Lemma`, `UPosTag`, `XPosTag`, `head.i`.

- `DepRel`: indicates the syntactic dependence relationship of the word to the main word in the sentence.
- `Form`: represents the word's surface form and how it appears in the text.

[1] https://spaCy.io/models/it#it_core_news_lg

[2] https://github.com/explosion/spacy-models/releases/tag/it_core_news_lg-3.7.0

Table 1: Comparison of the performance metrics of the three models across the entire dataset.

	Accuracy	Recall	Precision	F1 Score	Benchmark Match
P-based	0.70	0.79	0.87	0.83	78.90%
S-based	0.67	0.86	0.75	0.80	85.88%
Hybrid	0.67	0.90	0.73	0.80	90.20%

Table 2: Comparison of performance metrics of the three models concerning modifier adjectives.

	Accuracy	Recall	Precision	F1 Score	Benchmark Match
P-based	0.76	0.83	0.90	0.87	83.43%
S-based	0.68	0.88	0.75	0.81	88.25%
Hybrid	0.70	0.93	0.73	0.82	93.37%

Table 3: Comparison of performance metrics of the three models concerning verb-object combination.

	Accuracy	Recall	Precision	F1 Score	Benchmark Match
P-based	0.63	0.73	0.82	0.77	73.33%
S-based	0.66	0.83	0.76	0.79	82.96%
Hybrid	0.64	0.86	0.71	0.78	86.30%

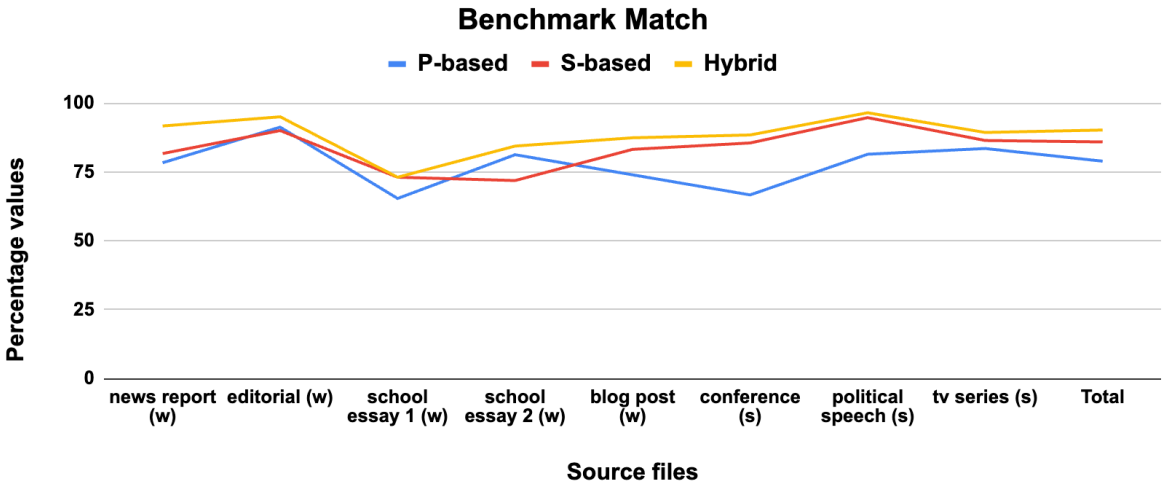

Figure 1: Benchmark Match values per file related to the entire dataset (w=written, s=speech).

- `Lemma`: is the basic form of a word that appears in dictionaries.
- `UPosTag` (Universal Part of Speech Tag): indicates the grammatical category of the word according to the universal POS tag scheme.
- `XPosTag` (Extended Part of Speech Tag): provides an extended POS tag that can include additional information.
- `head.i` (Head index): indicates the index of the word to which the current word is directly connected as a child in the sentence tree structure.

This information alone is not sufficient to fully understand the sentence's logical structure. Therefore, we identified several syntactic rules translated into Python functions to check the currently examined word and its head and determine whether it is part of an amode or Vdobj word combination. These rules were crucial in increasing the model's accuracy and precision, by cross-using the values of the different linguistic information provided by the parsing output. Writing these rules is particularly complex, as Italian is a morphologically and syntactically rich language with relatively free word order. For this reason, we proceeded step by step by analyzing the results obtained from time to time

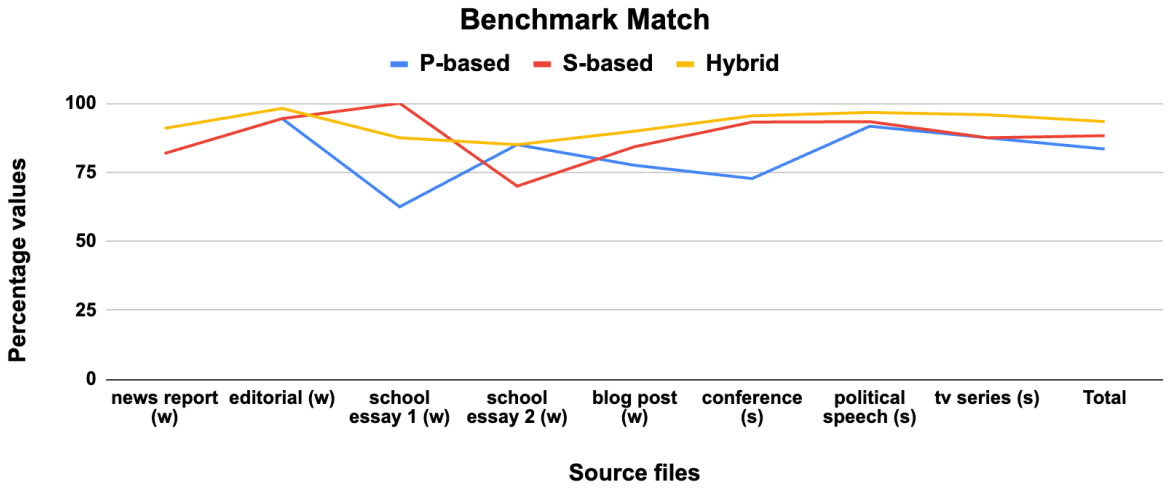

Figure 2: Benchmark Match values per file related to the modifier adjectives (w=written, s=spoken).

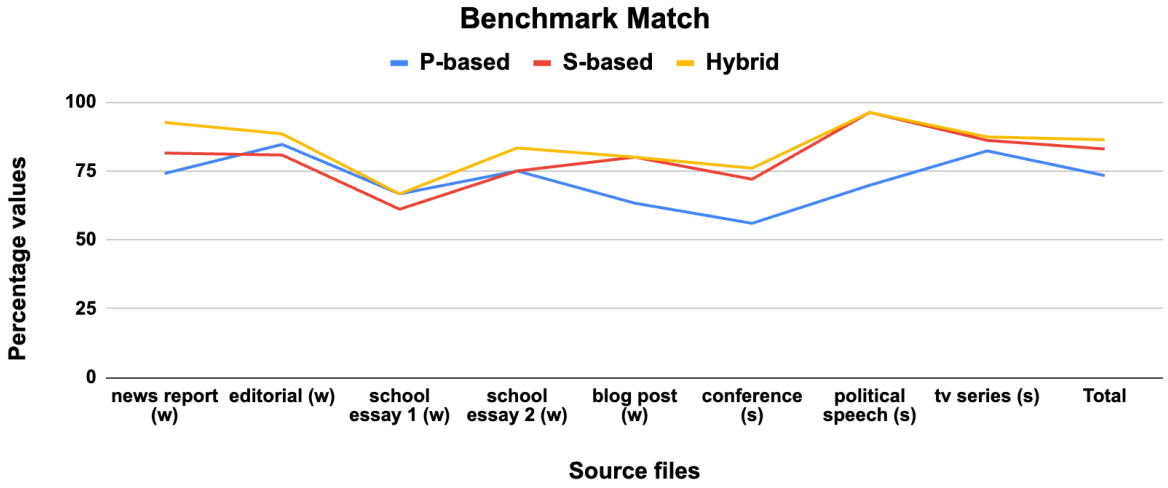

Figure 3: Benchmark Match values per file related to the verb-object combination (w=written, s=spoken).

and checking for incorrectly classified words to add rules, allowing the model to identify as many word combinations as possible. It is important to emphasize that the Python rules are specifically designed for the Italian language.

Some of the most important grammar rules that have been translated into Python code are now given. The first function recognizes a direct verbal object (Vdobj) with the obj relation with root as the dependency, while simultaneously verifying that the UPosTag of the root is VERB.

```
if token.dep_ == "obj" and
   token.head.dep_ == "ROOT"
   and token.pos_=="NOUN" and
   token.head.pos_ == "VERB"
```

This rule is able to recognize the combination of words *hanno fama* in Example 5.

Example 5. Molto note per le proprietà minerali delle acque sono le sorgenti di nitrodi e di olmitello, le loro virtù terapeutiche **hanno fama** mondiale. *Well-known for the mineral properties of the waters are the nitrodi and holmitello springs, their therapeutic virtues are world-renowned.*

Conversely, the function below is designed to identify AMOD when the 'amod' relation exists, with 'obj' as the dependency, and the UPosTag of the 'obj' token is NOUN.

```
if token.dep_ == "amod" and
   token.head.dep_ == "obj"
   and token.pos_=="ADJ" and
   token.head.pos_ == "NOUN"
```

The previous rule is able to recognize the word combination *straordinarie proprietà* in Example 6.

Example 6. Poi arrivarono i romani e scoprirono le **straordinarie proprietà** delle acque calde.
Then the Romans came and discovered the extraordinary properties of hot water.

In total, we created 18 functions to help us in identifying amod and Vdobj syntactic patterns. These functions were subsequently added to a function array. Each word was parsed from the function array, and upon finding a match, the result was saved in our data structure.

```
for token in line:
  for fun in functionsList:
    if fun(token):
      found="*"
```

At the end of this step, we obtained a data structure without duplicates of all word combinations categorized as amod or Vdobj, which was used as the input for the next step.

3.4.3. Statistical analysis of the model

The performance of the three approaches (P-based, S-based and Hybrid) was compared and evaluated through the usual measures of accuracy, precision, recall, F1 score. We defined in addition the *benchmark match*, which represents the percentage between the predictions generated by the model and the corresponding class labels in the benchmark file. It indicates how well the model aligns with the correct predictions established by the benchmark file, demonstrating its reliability and consistency against a validation dataset. The formula is $bm = 100 * (TP + TN)/(TP + TN + FN)$, where TP=True Positive, TN=True Negative, and FN=False Negative.

The Hybrid approach outperforms the P- and S-based approaches for the benchmark match and for recall. This better performance is observable across the entire dataset (Table 1), as well as for each of the syntactic relations taken individually (Tables 2 and 3). For the amod relation, the Hybrid approach reaches 93,37% of the benchmark match. This score can be regarded as highly positive in the context of candidate collocation identification. As expected, the P-based approach has better precision and worse recall, suggesting it has the lowest number of false positives but a reduced ability to identify positive instances. Conversely, the S-based approach shows low precision and high recall. It is worth noting that all the three methods have poorer results in detecting Vdobj relations compared to amod relations (Table 3), as in Vdobj relations the two words can be distant and in inverted order. However, the P-based approach is the one that has the most significant loss in benchmark match for Vdobj combinations (-10% compared to the amod relation).

In Figure 1, the benchmark match values related to the three approaches and the entire dataset are plotted as a function of the single sample files. Similar information is shown in Figure 2 about amod relation alone and in Figure 3 about Vdobj combinations alone. The figures allow for an evaluation of possible register influences on detection accuracy. The texts where the three approaches exhibit the most significant differences are two spoken texts, with a relatively formal register: the conference and the political speech, where the P-based approach has the worst results (Figure 1).

Overall, the Hybrid model validates our predictions and aligns more closely with the correct predictions established by the benchmark set, proving its reliability in complying with the gold standard of human annotation. The benefit of integrating the positional part of-speech and syntactic information for candidate collocation extraction is thus confirmed.

4. Conclusions and future work

Focusing on the automatic identification of candidate collocations in Italian corpora for lexicographic purposes, this study reports on an experiment aimed at comparing and evaluating the two most commonly used candidate detection approaches - the P-based and the S-based approach - with a third hybrid method resulting from the integration of the two previous ones. The evaluation of this step is crucial in order to assess the quality of candidate collocations with respect to specific criteria: their grammatical well-formedness (Seretan, 2011). Our assumption was that this quality would benefit from the integration of robust regex-over-pos methods with syntax-based approaches, despite the challenges posed by parsing large amounts of text in a morphosyntactically rich language like Italian. Results show that the Hybrid approach outperforms the two other methods in benchmark match and recall values, confirming the validity of our assumptions. Further work is still needed to optimise the model as precision, accuracy and F1 score obtain higher values with a P-based approach. By implementing additional Python rules, e.g. negative rules (i.e. rules capable of removing false positives) we believe we can enhance the performance of the S-based approach by refining the predictive accuracy while reducing false positives. This, when combined with the outcomes of the P-based approach, is expected to result in an overall enhancement in the model's performance.

Although the robustness of post-tagging can bal-

ance to some extent the lower accuracy of syntactic parsing, the rules applied in detecting syntactic relations after parsing need refinements to reduce errors resulting from false positives. One limitation of this experiment derives from using only two syntactic relations, whereas the final procedure for dictionary entry selection will need to consider a larger set of relations. However, the conclusion that can be drawn is that pursuing a hybrid approach to candidate collocation identification is worthwhile, as it leads to an improvement in the quality of results.

5. Acknowledgements

The research has been funded by the Italian Ministry of Research (MUR), PRIN: Research Projects of Major National Interest – Call 2022 - Prot. 2022HXZR5E. The title of the project is: *DICI-A: A Learner Dictionary of Italian Collocations*.

6. References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Giuseppe Attardi, Simone Saletti, and Maria Simi. 2015. Evolution of italian treebank and dependency parsing towards universal dependencies. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015 - Trento - 3-4 December 2015*, Torino. Accademia University Press.

Priscilla Benedetti, Damiano Perri, Marco Simonetti, Osvaldo Gervasi, Gianluca Reali, and Mauro Femminella. 2020. Skin cancer classification using inception network and transfer learning. *Lecture Notes in Computer Science*, 12249 LNCS:536 – 545. Green Open Access.

Morton Benson, Evelyn Benson, and Robert Ilson. 1986. *The BBI Dictionary of English Word Combinations*. Benjamins, Amsterdam.

Elisabeth Breidt. 1993. Extraction of v-n-collocations from text corpora: A feasibility study for german. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 74–83, Columbus, USA.

Sara Castagnoli, Gianluca E. Lebani, Alessandro Lenci, Francesca Masini, Malvina Nissim, and Lucia C. Passaro. 2016. Pos-patterns or syntax? comparing methods for extracting word combinations. In Gloria Corpas Pastor, editor, *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, pages 116–128. Tradulex, Geneve.

Yaacov Choueka. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *User-Oriented Content-Based Text and Image Handling*, page 609–623, Paris, FRA. Le Centre de Hautes Etudes Internationales d'informatique Documentaire.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Béatrice Daille. 1994. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Ph.D. thesis, Université Paris 7.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

Patrick Hanks. 2012. Corpus evidence and electronic lexicography. In Sylviane Granger and Magali Paquot, editors, *Electronic Lexicography*, pages 57—-82. Oxford University Press, Oxford.

Andrew Hardie. 2012. Cqpweb combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Brigitte Krenn. 2000. Collocation mining: Exploiting corpora for collocation idenfication and representation. In *Proceedings of KONVENS 2000*, Ilmenau, Germany.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317—-324, Morristown, NJ, USA.

Vincenzo Lo Cascio. 2013. *Dizionario Combinatorio Italiano*. Benjamins, Amsterdam.

Yajuan Lü and M. Zhou. 2004. Collocation translation acquisition using monolingual corpora. In *Annual Meeting of the Association for Computational Linguistics*.

Colin McIntosh, Ben Francis, and Richard Poole. 2002. *Oxford Collocations Dictionary for Students of English*. Oxford University Press, Oxford.

Alfredo Milani, Valentina Franzoni, and Giulio Biondi. 2021. Parsing tools for italian phraseological units. In *Computational Science and Its Applications – ICCSA 2021*, pages 427–435, Cham. Springer International Publishing.

Brigitte Orliac and Mike Dillinger. 2003. Collocation extraction for machine translation. In *Proceedings of Machine Translation Summit IX: Papers*, New Orleans, USA.

Magali Paquot. 2015. Lexicographyand phraseology. In Douglas Biber and Randi Reppen, editors, *The Cambridge Handbook of English Corpus Linguistics*, Cambridge Handbooks in Language and Linguistics, pages 460–477. Cambridge University Press.

Damiano Perri, Marco Simonetti, and Osvaldo Gervasi. 2022. Synthetic data generation to speed-up the object recognition pipeline. *Electronics (Switzerland)*, 11(1). Gold Open Access.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Julia Ritz. 2006. Collocation extraction: Needs, feeds and results of an extraction system for german. In *Proceedings of the workshop on Multiword-expressions in a multilingual context at the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 41–48, Trento, Italy.

Michael Rundell. 2010. *Macmillan Collocations Dictionary for learners of English*. Macmillan Education, London.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

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

Tianze Shi and Lillian Lee. 2020. Extracting headless mwes from dependency parse trees: Parsing, tagging, and joint modeling approaches. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8780–8794, Online. Association for Computational Linguistics.

Katalin Ilona Simkó, Viktória Kovács, and Veronika Vincze. 2017. Uszeged: Identifying verbal multiword expressions with pos tagging and parsing techniques. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 48–53, Valencia, Spain. Association for Computational Linguistics.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143—-177.

Stefania Spina. 2014. Il perugia corpus: una risorsa di riferimento per l'italiano. composizione, annotazione e valutazione. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014*, volume 1, pages 354–359, Pisa. Pisa University Press.

Stefania Spina. 2016. Learner corpus research and phraseology in italian as a second language: The case of the dici-a, a learner dictionary of italian collocations. In Begoña Sanromán Vilas, editor, *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching*, pages 219–244. Memoires de la Societe Neophilologique de Helsinki, Helsinky.

Stefania Spina. 2020. The role of learner corpus research in the study of l2 phraseology: main contributions and future directions. *Rivista di psicolinguistica applicata - Journal of Applied Psycholinguistics*, XX(2):35–52.

Paola Tiberii. 2012. *Dizionario delle collocazioni*. Zanichelli, Bologna.

Francesco Urzì. 2009. *Dizionario delle Combinazioni Lessicali*. Convivium, Luxemburg.

Hau Wu and Ming Zhou. 2003. Synonymous collocation extraction using translation information. In *Proceeding of the Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 120—-127.