# Towards a Statistical-Enriched Corpus Containing Portuguese Collocations in Use: Reviewing Possible Extraction Tools

Ângela Costa[1,3(✉)] and Luísa Coheur[1,2]

[1] INESC-ID Lisboa, Lisbon, Portugal
{angela,luisa.coheur}@l2f.inesc-id.pt
[2] Instituto Superior Técnico, Lisbon, Portugal
[3] Centro de Linguística da, Universidade Nova de Lisboa, Lisbon, Portugal

**Abstract.** Collocations are a main problem for any natural language processing task, from machine translation to summarization. With the goal of building a corpus with collocations, enriched with statistical information about them, we survey, in this paper, four tools for extracting collocations. These tools allow us to collect sentences with collocations, and also to gather statistics on this particular type of co-ocurrences, like Mutual Information and Log likelihood values.

**Keywords:** Collocations · Wortschatz · DeepDict · CRPC · Sketch engine

## 1 Introduction

Collocations, here understood as "a privileged lexical co-ocurrence of two or more linguistic elements that establish between themselves a syntactic relation." [10], are a type of multiword expressions that have always caused difficulties to natural language processing tasks. Several approaches have been proposed to retrieve collocations from the analysis of large samples of textual data. These techniques automatically produce large numbers of collocations along with statistical figures intended to reflect the relevance of the associations. Some work has been done on the evaluation of these techniques. For instance, the work described in [1], working only with English, evaluated WordSmith Tools 4, Collocate, Xaira and Ngram Statistics Package, regarding the capacity to extract collocations without keywords, measures of association, capacity to handle xml files, extract multiword collocations, handle multiple files at the same time, and the presence of a Graphical User Interface. The work from [4] also focused on an evaluation of extraction tools, for Portuguese, using several technical parameters like, for instance, how the data are organized and stored. As a result of using extraction techniques, [8] extracted and manually evaluated possible candidates and created a large list of multiword expressions in Portuguese.


J. Silva et al. (Eds.): PROPOR 2016, LNAI 9727, pp. 319–329, 2016.
DOI: 10.1007/978-3-319-41552-9_32

In this paper, having in mind the construction of a statistical-enriched corpus containing Portuguese collocations in use, we survey four online tools that allow to automatically find collocation candidates: Wortschatz[1], DeepDict[2], Reference Corpus of Contemporary Portuguese[3] and Sketch Engine[4]. The evaluation of these tools allow us to understand which one better fits our purpose of creating a corpus of collocations in context that is enriched with statistical information. There are other tools to do queries on Portuguese corpora, like Cetempublico or Corpus do Português but we found the aforementioned ones more suitable for the extraction of collocations in context and with statistical information.

In Sect. 2, we present our methodology. The extraction tools are described in Sect. 3. In Sect. 4, we compare the tools. Finally, in Sect. 5, we highlight the main conclusions and point to future directions.

## 2 Methodology

In this section we present the methodology followed to collect collocations in context.

As frequency is not, by itself, a defining trait of the collocations, we decided to use nouns with high and low frequency in the language taken from a Portuguese corpus of frequency available in COMPARA[5]. We used a total of 10 nouns: 5 with high frequencies, and 5 with low frequencies. Then, we submitted these nouns to the surveyed tools. Each noun is being tested as the base of a collocation, and we target to find possible collocates.

Each one of the tools is then evaluated (description in Sect. 4.2), according to the parameters that we will now describe below, which we considered to be the most important regarding our research purposes.

Starting with the **processing time**, we want it to be fast, as we need to perform several searches. Having the option to establish a **frequency threshold** is also very important, so that we can find the most salient collocations, and not simply common combinations of words. The search should be done by **lemma**, because only searching one word form can produce limited results. Determine the **window span** is also relevant. For instance, we want to capture the collocation *strong tea* in a wider span, like in *the tea was very strong*. As using **syntactic patterns** has proven to be a successful way to extract collocations, we would like to have this option available. Moreover, we would like to use a tool that **distinguishes the varieties of Portuguese** and that **does not have repeated texts in its corpora**, in order to ensure the representativeness of the extracted collocations and statistics. It is also an advantage to be able to **analyse the collocations in their context**, rather than have access to a list of co-occurring words and a score. Also, we would like to be able to **download these sentences**,

[1] http://corpora.informatik.uni-leipzig.de.
[2] https://gramtrans.com/deepdict/.
[3] http://www.clul.ul.pt/pt/recursos/.
[4] https://www.sketchengine.co.uk.
[5] http://www.linguateca.pt/COMPARA/listas_freq.php.

and **upload different corpora** for analysis. Regarding metrics, we would like to have information about **Mutual Information, Log-likelihood and Dice coefficient**. Also, we would prefer to use a **freely available** tool. Finally, we will also compare the number of correct collocations that were selected by the systems (Sect. 4.1).

## 3 The Surveyed Tools

In this section, we present the tools that we have analyzed: Wortschatz, DeepDict, Reference Corpus of Contemporary Portuguese, and Sketch Engine. The word *momento* will be used for illustration proposes.

### 3.1 Wortschatz

Wortschatz is a system developed by the University of Leipzig that uses the Leipzig Corpora Collection. This corpus exists for more than 250 languages, all data is searchable. The Portuguese data consists of a newspaper corpus based on material crawled in 2011 and the Wikipedia also collected in 2011. For the European variety of Portuguese, there are 2.540.587 sentences and 53.879.750 tokens. The Brazilian variety has 25.008.883 sentences and 486.724.987 tokens. The variety of Macau has 392,371 sentences and 8.672.381 tokens. The search can be done in one of the varieties or all. Several tools are used for preprocessing the corpus (for instance, tokenization, word frequency calculation, word co-occurrence calculation). There are also post-processing tools, like POS tagging and lemmatization. All the tools are available for download from their website, as well as most of the corpora.

To do the search, typing the word is the only option, as Wortschatz is the only evaluated tool that does not allow to establish a frequency threshold, as well as perform the search by lemma. On Fig. 1, we can see the search result for the word *momento*. On the top, we can find the number of occurrences, the rank and frequency class and the sentences with the queried word. Bellow, the co-occurrences, the left neighbor co-occurrences and the right ones. If we click on one of the words that co-occurs, we can see the sentences where the candidate collocate occurs, but not the collocation. Finally, there is no download option.

### 3.2 DeepDict

DeepDict is a free tool that allows to build complex dictionary entries and context overviews for a given word on the fly. Word relations are based on Constraint Grammar[6] dependency analysis and grammatical functions, not just co-occurrence like Wortschatz. In the case of Portuguese, the multi-level Constraint Grammar parser used is PALAVRAS[7] and the corpus used is

[6] http://beta.visl.sdu.dk/constraint_grammar.html.
[7] http://linguateca.dei.uc.pt/Floresta/InicialFloresta.html.


**Fig. 1.** Wortschatz results for the noun *momento*.

Floresta Sintá(c)tica[8], which is a collection of sentences that have been morfos-syntactically analyzed, producing a treebank of 1.000.000 words collected from CETEM Publico [9].

To look up collocational candidates, we type the word, discriminate the word class and the language. There are also other advanced options available, like establishing a lexical frequency threshold (used to filter rare words), a minimum occurrence (rule out rare relations or include them) and a minimum relative frequency (set a threshold for co-occurrences). As for the window span, DeepDict as well as Wortschatz do not have that option.

Figure 2 shows how results are displayed, considering the word *momento*.

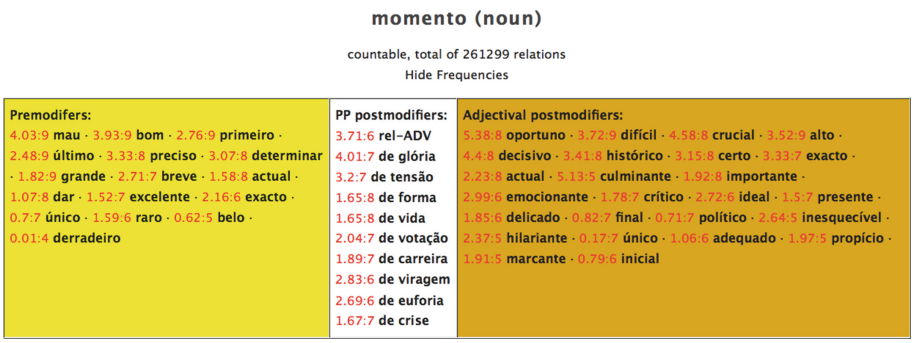



**Fig. 2.** DeepDict results for the noun *momento*. (Color figure online)

For each search, relative and absolute frequency values are provided for each relation (premodifiers, PP postmodifiers and Adjectival postmodifiers). Frequency values (in red) can be clicked to see detailed statistics. Considering the

[8] http://www.linguateca.pt/floresta/principal.html.

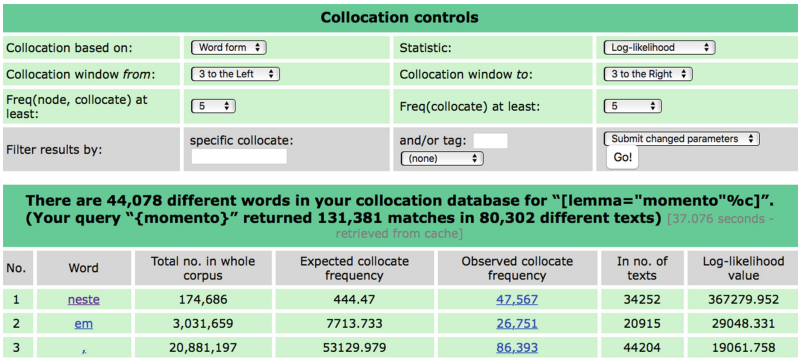
co-occurring word mau as example, 4.03 is the co-occurrence strength between the lookup word and the relative frequency, and 9 is the dual logarithmic value of absolute frequency (scale is from 0 to 9). Red numbers can be clicked to show examples in concordance form, if available, and the statistics of these forms. Like Wortschatz, these data are not available for download.

### 3.3 Reference Corpus of Contemporary Portuguese

The Reference Corpus of Contemporary Portuguese (CRPC) is an electronic corpus of European Portuguese and varieties (Brazil, Angola, Cape Verde, Guinea-Bissau, Mozambique, S. Tome and Principe, Goa, Macao and East-Timor). It contains 311.4 million words and covers several types of written texts (literary, newspaper, technical, etc.) and spoken texts. Searches can be made online on the written subpart of the corpus (309 M). The texts were tokenized using the LX tokenizer [3]. The part-of-speech tagging was trained based on a memory-based tagger [5]. Finally, the lemmatization was done with a Portuguese version of the MBLEM lemmatizer [2].

First, we have to select a corpus, either the European Portuguese or the whole corpus, including varieties. Then, we type a word in the search box (or syntactic pattern). We select the number of hits per page and we can also restrict our search to a specific corpus (laws, newspapers, school books, etc.). This is the only tool that allows this specific type of search. We are then presented with the search word in context, after which we can create a collocation database with the list of words that co-occur with the retrieved word pattern.

Figure 3 shows the results given the word *momento*.



**Collocation controls**

| | | | |
|---|---|---|---|
| Collocation based on: | Word form | Statistic: | Log-likelihood |
| Collocation window *from*: | 3 to the Left | Collocation window *to*: | 3 to the Right |
| Freq(node, collocate) at least: | 5 | Freq(collocate) at least: | 5 |
| Filter results by: | specific collocate: | and/or tag: (none) | Submit changed parameters, Go! |

**There are 44,078 different words in your collocation database for "[lemma="momento"%c]". (Your query "{momento}" returned 131,381 matches in 80,302 different texts)** [37.076 seconds - retrieved from cache]

| No. | Word | Total no. in whole corpus | Expected collocate frequency | Observed collocate frequency | In no. of texts | Log-likelihood value |
|---|---|---|---|---|---|---|
| 1 | neste | 174,686 | 444.47 | 47,567 | 34252 | 367279.952 |
| 2 | em | 3,031,659 | 7713.733 | 26,751 | 20915 | 29048.331 |
| 3 | , | 20,881,197 | 53129.979 | 86,393 | 44204 | 19061.758 |

**Fig. 3.** CRPC results for the noun *momento*.

We can decide the maximum window span to be used. On the collocation controls, unlike the previously described systems, we can change the statistical measure used and the frequency threshold. If we click on observed collocate frequency, we can see the collocation in context. Contrary to the other systems,

we can download these sentences and all the statistics. Regarding the processing time for the word *momento*, Wortschatz and DeepDict took less than 6 s, while CRPC took approximately 46 s.

### 3.4 Sketch Engine

Sketch engine is a paid corpus software interface that works online. It allows to see concordances for any word, phrase or grammatical construction; it also shows each word grammatical and collocational behavior. It has 200 corpora in 82 languages, but we can also upload our own corpus, being the only tool that presents this feature. There are available corpora for two varieties of Portuguese: European and Brazilian. The parser used, like for Deep Dict, is PALAVRAS. There are several valences to this software, like creating a Thesaurus, word lists or comparing occurrences of two words, but we will focus on the options that allow extracting collocations: "word sketch" and "concordance". They will be detailed in the next paragraphs.

**Word Sketch.** Selecting "word sketch", will allow to see a word's grammatical and collocational behavior. We type the lemma and specify its part-of-speech. On the advanced options, we choose the European Portuguese corpus (ptTenTen11 with 3.245.834.337 tokens from web pages). We can select a minimum frequency and/or score (score is defined as logDice[9]) and the minimum value of co-occurrence. We can also choose to cluster[10], or not, the collocates and decide the maximum number of items in these grammatical relations. Collocations can also be sorted by salience[11] or by raw frequency. The word sketch, in addition to using a well-founded salience statistic and lemmatization, uses grammar patterns. Rather than looking at an arbitrary window of text around the headword, each grammatical relation that the word participates in is taken into consideration. For Portuguese, there are 11 grammatical relations. The word sketch then provides one list of collocates for each grammatical relation the word participates in, and the results will be presented in syntactic relation clusters (object_of, subject_of or n_modifier). CRPC also allows a search by a syntactic pattern, for instance, noun + adjective or a word, like *momento* + adjective (we have not used this option for the present evaluation). DeepDict does not allow a search by pattern, but the results are presented in relation clusters.

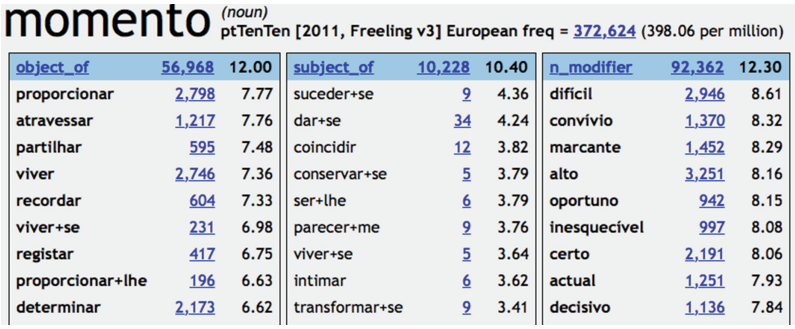
As usual, Fig. 4 shows the sketch for the word *momento*.

As mentioned before, the potential collocates are presented grouped according to the grammatical relation in which they occur. The first score in front of

[9] This measure is based only on a frequency of words w1 and w2 and bigram w1 w2, it is not affected by the size of the corpus.

[10] If the clustering option is selected, the collocates within a word sketch are clustered according to any such clusters from the distributional thesaurus that they appear in. The words from the thesaurus are clustered according to their distributional similarity scores.

[11] Salience is a statistical measure of how salient a word or lemma is in a given context, given the frequency of the word and the context. This is measured with logDice.



momento *(noun)*
ptTenTen [2011, Freeling v3] European freq = 372,624 (398.06 per million)

| object_of | 56,968 | 12.00 | subject_of | 10,228 | 10.40 | n_modifier | 92,362 | 12.30 |
|---|---|---|---|---|---|---|---|---|
| proporcionar | 2,798 | 7.77 | suceder+se | 9 | 4.36 | difícil | 2,946 | 8.61 |
| atravessar | 1,217 | 7.76 | dar+se | 34 | 4.24 | convívio | 1,370 | 8.32 |
| partilhar | 595 | 7.48 | coincidir | 12 | 3.82 | marcante | 1,452 | 8.29 |
| viver | 2,746 | 7.36 | conservar+se | 5 | 3.79 | alto | 3,251 | 8.16 |
| recordar | 604 | 7.33 | ser+lhe | 6 | 3.79 | oportuno | 942 | 8.15 |
| viver+se | 231 | 6.98 | parecer+me | 9 | 3.76 | inesquecível | 997 | 8.08 |
| registar | 417 | 6.75 | viver+se | 5 | 3.64 | certo | 2,191 | 8.06 |
| proporcionar+lhe | 196 | 6.63 | intimar | 6 | 3.62 | actual | 1,251 | 7.93 |
| determinar | 2,173 | 6.62 | transformar+se | 9 | 3.41 | decisivo | 1,136 | 7.84 |

**Fig. 4.** Word sketch results for the noun *momento*.

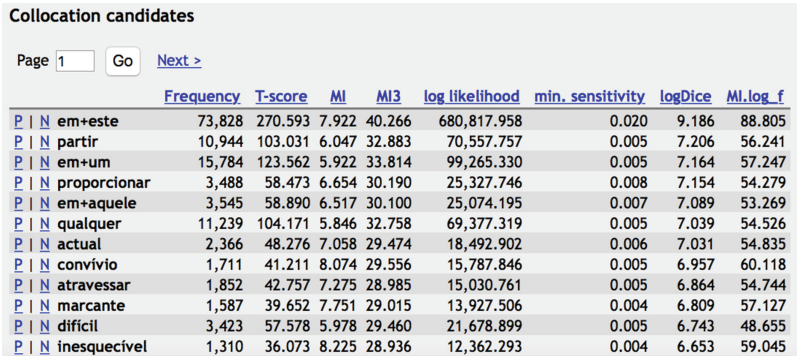
each collocate candidate is the word frequency. If we click on it, we can see the corpus contexts in which the node word and its collocate co-occur. The second number is the frequency within the cluster. The word sketches and the examples of the corpus can be downloaded in xml or txt.

**Concordance.** Another way to extract collocations, is to use the "concordance" option. We type the word, select the corpus, and attain all the occurrences of the searched word in the corpus. The next step is to build the collocation list. This is done by creating a list of words statistically associated with the words (node) in the query. In this search menu, we select the attribute (word, part of speech tag, lemma, etc.). We can specify the range (span of text) around our node word when considering candidates. The defaults are $-3$ (3 tokens before the node) and 3 (3 tokens after the node). And we can also specify thresholds on the frequency of the candidate in the whole corpus and the frequency within the range. We can stipulate which statistics are displayed, and the statistic to sort by. The statistics available are relative frequency, T-Score, MI-Score, MI3-Score, log-likelihood, minimum sensitivity, logDice and MI.log-f. More details on the metrics can be found in [6]. This is the extraction tool that provides more statistical information, although CRPC also gives a considerate amount of statistics.

Figure 5 shows the sketch for the word *momento*. To see the results ordered according to a metric, we just have to click on the chosen metric. We can see the word co-ocurrence on the corpus, if we click on "P". The metrics results and the corpora can be downloaded.

## 4 Comparison Between Systems

On the next sections, we will assess each tool according to the collocations they were able to find, the extracted corpora and statistics that resulted from that task.



Collocation candidates

Page 1 Go Next >

| | Frequency | T-score | MI | MI3 | log likelihood | min. sensitivity | logDice | MI.log_f |
|---|---|---|---|---|---|---|---|---|
| P \| N em+este | 73,828 | 270.593 | 7.922 | 40.266 | 680,817.958 | 0.020 | 9.186 | 88.805 |
| P \| N partir | 10,944 | 103.031 | 6.047 | 32.883 | 70,557.757 | 0.005 | 7.206 | 56.241 |
| P \| N em+um | 15,784 | 123.562 | 5.922 | 33.814 | 99,265.330 | 0.005 | 7.164 | 57.247 |
| P \| N proporcionar | 3,488 | 58.473 | 6.654 | 30.190 | 25,327.746 | 0.008 | 7.154 | 54.279 |
| P \| N em+aquele | 3,545 | 58.890 | 6.517 | 30.100 | 25,074.195 | 0.007 | 7.089 | 53.269 |
| P \| N qualquer | 11,239 | 104.171 | 5.846 | 32.758 | 69,377.319 | 0.005 | 7.039 | 54.526 |
| P \| N actual | 2,366 | 48.276 | 7.058 | 29.474 | 18,492.902 | 0.006 | 7.031 | 54.835 |
| P \| N convívio | 1,711 | 41.211 | 8.074 | 29.556 | 15,787.846 | 0.005 | 6.957 | 60.118 |
| P \| N atravessar | 1,852 | 42.757 | 7.275 | 28.985 | 15,030.761 | 0.005 | 6.864 | 54.744 |
| P \| N marcante | 1,587 | 39.652 | 7.751 | 29.015 | 13,927.506 | 0.004 | 6.809 | 57.127 |
| P \| N difícil | 3,423 | 57.578 | 5.978 | 29.460 | 21,678.899 | 0.005 | 6.743 | 48.655 |
| P \| N inesquecível | 1,310 | 36.073 | 8.225 | 28.936 | 12,362.293 | 0.004 | 6.653 | 59.045 |

**Fig. 5.** Concordance results for the noun *momento*.

### 4.1 Finding Collocations

As previously mentioned, we have done the searches, on each system, using 10 nouns as the base of the collocations. These nouns were selected from a list of frequent nouns in Portuguese. These nouns are part of the current vocabulary of Portuguese (opposed to specialized vocabulary) and, as tools corpora are mainly extracted from newspapers and the web, we believe there were no problems of low representativeness of those nouns.

For this paper, we have only analyzed the first 10 extractions from each search tool. The setup used, when allowed by the system, was a minimum frequency of 5 and a window span of 3 words to the right and left of the node. Table 1 shows the number of collocations extracted and validated by a linguist. The last column shows examples of collocations that were extracted by more than one system. As previously mentioned, the tools have different size internal corpora, which obviously skews the comparison. Still as only one of the tools allows the upload of a corpus, a comparison using the same texts would be impossible. That said, we will interpret the size of the corpora as an idiosyncrasy of each tool and the results should be interpreted in this light.

The system that was able to find more collocations was Word Sketch, showing that combining statistics and grammar patterns can be a successful way to extract collocations, rather than simply count occurrences and frequencies, like Wortschatz. Regarding the words that were selected but that were not collocates, we found among them articles (*o momento*), prepositions (*na verdade*) and words that usually occur in the same semantic field (*autógrafo* and *fotografia*). We should also point out that words that have a lower frequency score are the ones that show fewer candidate options and more restriction in the choice of the collocates (*diagnóstico precoce*, *escolaridade obrigatória*).

### 4.2 Collecting Corpora and Statistics

First of all, for the purpose of our research, we only wanted to use the European variety of Portuguese. All engines distinguish between varieties and DeepDict

**Table 1.** Comparison between Wortschatz (Woc), DeepDict (DD), CRPC, Word Sketch (SE-1), Concordance (SE-2) extractions.

| Word | WoC | DD | CRPC | SE-1 | SE-2 | Examples |
|---|---|---|---|---|---|---|
| momento | 0 | 5 | 1 | 2 | 1 | oportuno (2); decisivo (2) |
| fim | 0 | 1 | 1 | 4 | 3 | lucrativo (4); pôr (2) |
| verdade | 1 | 1 | 0 | 5 | 2 | absoluto (3) |
| certeza | 1 | 2 | 3 | 4 | 2 | absoluto (4); ter (4) |
| força | 1 | 0 | 0 | 3 | 3 | de vontade (3) |
| adversidade | 1 | 2 | 3 | 6 | 5 | superar (3); climatérico (3) |
| autógrafo | 0 | 0 | 2 | 6 | 2 | pedir (3); dar (2) |
| fumador | 2 | 4 | 2 | 3 | 2 | inveterado (5); passivo (5) |
| diagnóstico | 1 | 1 | 1 | 2 | 1 | precoce (5) |
| escolaridade | 1 | 1 | 2 | 1 | 2 | obrigatório (5) |
| **TOTAL** | **8** | **17** | **14** | **36** | **23** | |

only has available an European Portuguese Corpus. Its is also important that the corpora used are not very repetitive in its constitution because this can bias the statistical results, but we only spotted repeated sentences in the CRPC corpus. We also wanted to be able to see the collocations in the context of the corpus, but this criteria was only not accomplished by DeepDict that only shows the collocate in context and not the entire collocation. Apart from visualizing the collocations in context, we also wanted to be able to download the sentences where they occurred. In this case, only Sketch Engine and CRPC allow download of the data. Sketch Engine is the only one that allows the user to upload its own corpus and do searches on it.

The purpose of this evaluation was to access which tool better served the purpose of collecting a corpus enriched with collocations but also gather all the statistical information available. DeepDict gives us the relative frequency of a given relation and the absolute frequency in a scale from 0 to 9. By using the "word sketch" in Sketch Engine, we can obtain the word frequency and the frequency within the cluster. With the "concordance" option, the collocation candidates can be put in order according to several metrics: relative frequency, T-Score, MI-Score, MI3 -Score, log-likelihood, minimum sensitivity, logDice and MI.log-f. CRPC provides information on the total number of occurrences in the whole corpus, the expected collocate frequency, the observed collocate frequency and number of texts of the occurrence. Then we can also change the metrics and see the ranking of results according to the Mutual information, MI13, Z-score, T-score, Log-likelihood, Dice coefficient or simple rank by frequency. The results will change according to the chosen metric. Wortschatz only features the number of occurrences. Finally, we also took into consideration if the tool was free. From the four we have used, only Sketch Engine needs to be paid. DeepDict is paid for some languages, but not for Portuguese. Table 2 sums up the evaluation of all systems. Although Sketch Engine has two search options, here was presented as one as the results are very similar.

**Table 2.** Comparison between systems features.

| | Woc | DD | CRPC | SE |
|---|---|---|---|---|
| Processing time for the word *momento* | ≈ 1 s | ≈ 2 s | ≈ 46 s | ≈ 6 s |
| Frequency threshold | ✘ | ✓ | ✓ | ✓ |
| Search by lemma | ✘ | ✓ | ✓ | ✓ |
| Window span | ✘ | ✘ | ✓ | ✓ |
| Search by pattern | ✘ | ✘ | ✓ | ✓ |
| Distinguish varieties | ✓ | ✓ | ✓ | ✓ |
| No duplications in the corpus | ✓ | ✓ | ✘ | ✓ |
| Show co-ocurrences in the frpus | ✘ | ✓ | ✓ | ✓ |
| Download results | ✘ | ✘ | ✓ | ✓ |
| Upload corpus | ✘ | ✘ | ✘ | ✓ |
| Use several metrics | ✘ | ✘ | ✓ | ✓ |
| Free | ✓ | ✓ | ✓ | ✘ |

## 5 Conclusions

In this paper we aimed at doing an assessment of four extracting tools: DeepDict, Sketch Engine, CRPC and Wortschatz. We started by describing each systems functionalities, then we used them to extract collocations, having as a base 10 words selected from the corpus of frequencies. We evaluated the resulting corpus and the static information provided by each tool. Based on the mentioned steps, we conclude that Sketch Engine, despite not being free, shows great potencial as collocation extractor, being able to find more correct collocations than the other systems. Of course this result is influenced by the size of the corpus that Sketch Engine uses. This tool outperforms the others in several aspects of our evaluation, for instance, allowing a fine-grained tuning of the search setup (frequency threshold, window span, search by lemma or pattern). The Portuguese corpus used is divided into BP and EP and, from our observations so far, does not seem to have repeated sentences. The results can be presented according to different metrics, allowing to understand which one better suits the research purpose[12]. Finally, the co-occurrences are shown in context, the results can be downloaded and different corpora can be uploaded to the tool. If using a paying tool is not an option, CRPC also shows good results, only having the disadvantage of the repetitions in the corpus. In the future, the aforementioned methodology used here for evaluation purposes, will be used to build a larger statistically-enriched corpus containing Portuguese collocations in use that will allow us to understand more about their contexts and will help us build rules to extract them.

[12] [7] suggests that MI is generally used for a lexicographical purpose, while MI3 is probably more useful for second language learning.

**Acknowledgments.** The work was partially supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, reference UID/CEC/50021/2013. Ângela Costa is supported by PhD fellowship from FCT (SFRH/BD/85737/2012).

## References

1. Anagnostou, N.K., Weir, G.R.S.: Review of software applications for derivingcollocations. In: ICT in the Analysis, Teaching and Learning of Languages, Preprints of the ICTATLL Workshop 2006, Glasgow, pp. 91–100 (2006)
2. van den Bosch, A., Daelemans, W.: Memory-based morphological analysis. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, ACL 1999, pp. 285–292. Association for Computational Linguistics, Stroudsburg (1999). http://dx.org/10.3115/1034678.1034726
3. Branco, A., Silva, J.: Evaluating solutions for the rapid development of state-of-the-art pos taggers for Portuguese. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004). European Language Resources Association (ELRA), Lisbon. http://www.lrec-conf.org/proceedings/lrec2004/pdf/572.pdf, aCL Anthology Identifier: L04–1354
4. Correia, J.M.P.: Syntax Deep Explorer. Ph.D. thesis. Instituto Superior Técnico (2015)
5. Daelemans, W., Zavrel, J., Berck, P., Gillis, S.: Mbt: a memory-based part of speech tagger-generator. In: Proceedings of Fourth Workshop on Very Large Corpora, pp. 14–27. ACL SIGDAT (1996)
6. Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D.: The sketch engine. In: Proceedings of EURALEX (2004)
7. McEnery, T., Xiao, R., Tono, Y.: Corpus-Based Language Studies: An Advanced Resource Book. Taylor & Francis (2006)
8. Mendes, A., Antunes, S., do Nascimento, M.F.B., Miguel, J., Casteleiro, L.P., Sá, T.: Combina-pt: a large corpus-extracted and hand-checked lexical database of Portuguese multiword expressions. In: Proceedings of LREC, pp. 1900–1905 (2006)
9. Santos, D., Rocha, P.: Evaluating CETEMPúblico, a free resource for Portuguese. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pp. 450–457. Association for Computational Linguistics (2001)
10. Tutin, A., Grossmann, F.: Collocations régulières et irrégulières: esquisse de typologie du phénomène collocatif. Revue française de linguistique appliquée **7**(1), 7–25 (2002)