Extrinsic Corpus Evaluation with a Collocation Dictionary Task

Adam Kilgarriff,[1] **Miloš Jakubíček,**[1,2] **Vojtěch Kovář,**[1,2] **Pavel Rychlý,**[1,2]
Vít Baisa,[1,2] **Lucia Kocincová** [2]
[1] Lexical Computing, Ltd., 71, Freshfield Road, Brighton, UK
[2] NLP Centre, Faculty of Informatics, Masaryk University, Botanická 68a, Brno, Czech Republic
adam.kilgarriff@sketchengine.co.uk, {jak, xkovar3, pary, xbaisa, xkocinc}@fi.muni.cz

Abstract

The NLP researcher or application-builder often wonders "what corpus should I use, or should I build one of my own? If I build one of my own, how will I know if I have done a good job?" Currently there is very little help available for them. They are in need of a framework for evaluating corpora. We develop such a framework, in relation to corpora which aim for good coverage of 'general language'. The task we set is automatic creation of a publication-quality collocations dictionary. For a sample of 100 headwords of Czech and 100 of English, we identify a gold standard dataset of (ideally) all the collocations that should appear for these headwords in such a dictionary. The datasets are being made available alongside this paper. We then use them to determine precision and recall for a range of corpora, with a range of parameters.

Keywords: corpus, evaluation, collocation

1. Cooks and Farmers

Let us talk about food. Cooks prepare the food. It is their skill and ingenuity, their methods and strategies, their inspiration and imagination, that gives rise to delicacies rare and fine, to tickle the palate and delight the senses.

But any cook will say, take care with your ingredients! Make sure the fishes' eyes sparkle, the tomatoes are plump and firm, the peaches ripe with a rosy hue. No-one can prepare a first-rate meal from third-rate ingredients. Be aware of your sources, you want your produce to be from a farmer who cares about quality.

So it is with language technology. Those writing the applications are the cooks, those preparing corpora, the farmers. The applications are crucial to the success of the enterprise. But — and this becomes inescapable as more and more methods are based on learning from data — so too is the quality of the data they are based on.

2. Introduction

It is twenty years now since the field awoke to the importance of evaluation (Gaizauskas, 1998). This has usually been evaluation of systems, with the playing field leveled by all systems using the same data.

It has also been two decades since the merits of approaches based on data have been explored in earnest. Despite it being the same two decades in each case — and the near-tautology that better input will result in better output — the field still has nothing to say about how to evaluate a corpus. In the 1990s this could be justified by the lack of corpora: when there was only one corpus available, the question 'how good is it' was not worth asking. Also the base considerations of having data available in a tractable format, with characters correctly encoded and data distinct from metadata, have been priorities, with initiatives such as TEI and many projects described at the LREC conferences focusing on these questions. Corpora have been validated, though not evaluated. But now we are in an age of corpora on demand. The web provides a near boundless supply of text for a great many languages and text types, so it is easy to make a corpus that may well be good for a particular task. We need methods for evaluating.

In this paper we present a method for evaluating corpora and its implementation for Czech and English.

3. The Collocation Dictionary Creation Task

A corpus being good is relative to a task: different corpora will be good for different tasks. Many recent initiatives in language technology evaluation pay heed to this base truth with evaluations based on particular use cases.

Still, a well-designed evaluation can be relevant to a broad range of tasks, for two clusters of reasons:

- for most tasks, some criteria hold true
  - duplication of content is bad
  - junk (including "word salad", material in a computer language, material in the wrong human language) is bad
  - bigger is, all else being equal, better
- many tasks relate to "the language in general"
  - NLP tools such as POS-taggers and parsers are often built for "the language in general"
  - dictionaries and lexicons are typically for "the language in general".

(Kilgarriff et al., 2010) used the task of creating a collocations dictionary to evaluate word sketches (Kilgarriff et al., 2004). The method was to ask, for each of the twenty highest-scoring collocations for a sample of headwords, 'should this collocation be in a published collocations dictionary?' The Oxford Collocations Dictionary (Crowther et al., 2002, OCD) was taken as a reference point for such a dictionary. The evaluation was carried out for four languages. The word sketch evaluation was a variant of the

series of collocation-extraction evaluation exercises undertaken for German (Krenn et al., 2001) and others since.
A feature that all these evaluations share is that the same gold-standard data can be used to evaluate a number of components. The components are, in outline, the corpus, the NLP tools, and the statistic used to score and rank collocations. If we know the collocations that should have been delivered, then we can ask, as Krenn and Evert did, which statistic gives us the best result? Or we can ask, if we hold the corpus and statistics constant, which NLP tools give us the best result? Or, holding all else constant, which corpus is best?

4. Task Definition

The introduction to the Oxford Collocations Dictionary (OCD) states

> Collocation is the way words combine in a language to produce natural-sounding speech and writing. ... Combinations of words in a language can be ranged on a cline from the totally free — *see a man/car/book* — to the totally fixed and idiomatic — *not see the wood for the trees*. ... All these combinations, apart form those at the very extremes of the cline, can be called collocation. And it is combinations such as these — particularly in the 'medium-strength' area — that are vital to communicative competence in English. (Crowther et al., 2002, vii)

The 'extremes of the cline' are not in general high-frequency items, and this account of collocation fits well with corpus methods. We adopt it.[1]
Several further questions arose in the task definition:
Names In both Czech and English names and name-like items are usually capitalised. After some discussion, about *Hell's Angels* amongst others, we followed the most straightforward route: all capitalised items (also items including hyphens, numbers or other non-letters) to be excluded.
Recall The evaluation in (Kilgarriff et al., 2010) evaluated only precision, and there was no counterweight to helpfully-inclined evaluators being generous in accepting the proposed collocations. To compare one corpus to another, it is not enough to know which, of a limited set of candidates, are good: we need to know **all** the good ones.
As in comparable exercises in Information Retrieval (IR), asking human judges to judge all possible candidates is not economically viable. We adopt the 'pooling method', as used in IR exercises such as TREC:[2] find all candidates, according to a range of systems and parameters, to build a large set of candidates, which, we hope, includes all good items. The judge then judges those items.

[1] OCD policy is in contrast to the Macmillan Collocations Dictionary (Rundell, 2010), which takes a narrower and more focussed view of what to include, with 'likely to present a challenge to language learners' as central.

[2] See `trec.nist.gov`, in particular `trec.nist.gov/presentations/TREC5/15.html`

Grammar We represent a collocation by a lemma associated (unordered) with the headword. The headword has a word class associated with it (so that we can structure the sample by word class) but after some consideration we decided the collocating word should not. Also, although many collocation-finding systems use grammar, and, as part of their processing, identify the grammatical relation holding between collocates, we decided not to include grammatical relations in the representation. In both cases the reason was to minimise the dependency of the gold-standard dataset that we were producing, on particular accounts and vocabularies of grammatical relations or word classes, which would make it hard to use with a system that used a different vocabulary. We do assume lemmatisation — the mapping from inflected forms to dictionary headwords — and this causes some problems in English in relation to -ing and -ed verb forms[3] and Czech in relation to e.g. ne- adjective forms (*nemocný* – ill and *mocný* – powerfull).
It is a consequence of this decision that, if the headword is $hair$ (noun) we consider "brushing her hair" and "his hair brush" to be instances of the same collocation.
Grammar words, collocations of more than two words We needed to decide how to handle items such as *look at*, *on fire*, *criticise on the grounds that*, *male chauvinist pig*. The first two items are in the area in which the concept of collocation blends into that of grammatical patterning. This was not our core concern and would have raised many further questions. We took the pragmatic solution of a stoplist of grammar words. Combinations of headword + stoplist word would not count. This also meant that many collocations of three or more words resolved to just two non-stoplist words, e.g., ⟨*criticise, ground*⟩.
Beyond that, we pay no special attention to collocations of more than three words, so we have the three collocations ⟨*male, chauvinist*⟩, ⟨*male, pig*⟩ and ⟨*chauvinist, pig*⟩. There are far fewer three-and-more-content-word-collocations than two-word ones, so we did not expect the anomalies that this treatment might cause for the scoring scheme, to have any appreciable impact.
While we make no claims that the stopword list is an elegant solution, it is a transparent and easily-understood one. The stopword lists are published along with the gold standard data.
The question of what we represent as a collocation, is separate to the question of what we show to the judge. We show the judge the commonest form of the collocation (identified using the algorithm presented in (Kilgarriff et al., 2012)); whatever the collocation, we show the judge *male chauvinist pig* together with the collocation.
By using a gold standard dataset comprising lexical data (collocations) rather than corpus data (correct annotations on a text) we are evaluating generalisations drawn from the whole corpus. Each expert judgement is more informative than in the case where the expert judges corpus instances. Since a larger dataset will allow us better to distinguish signal from noise, the method will favour quantity – but not if too much quality is lost.

[3] It is often a judgement call whether or not the form is a gerundive noun or adjective in its own right, or should be treated as an inflected form on the verb.

The question we are asking the judges, "should this word be in a collocation dictionary" is a reasonable one, even if there are many judgement calls: it must be a reasonable one, since collocations dictionaries exist.

5. Creating Gold Standard Data

5.1. Sampling

Collocation dictionaries are for the core of the vocabulary: not the very rare words, or the grammatical words, but the common nouns, verbs and adjectives that make up 99% of the headword list in a standard dictionary.[4] OCD has collocations for 9000 headwords, but that seems a modest number. Intermediate-level learners' dictionaries typically have around 30,000 headwords.

We take a sample from the 30,000 commonest words, with the sample structured as in Table 1, nouns, verbs and adjectives in ratios of 3:2:2 and equal numbers for each frequency band. Within these constraints, the sampling was random. Table 1 also shows the words selected for English.

5.2. Finding Candidate Collocations

We now wished to prepare a set of collocation candidates, from both corpora and dictionaries, to present to our judges for them to say 'yes' or 'no' to. A key question was, how long should these lists be? Too long, and the cost was too great: too short, and our claim to be able to assess recall weakens. We decided on 500 for high-frequency words, 250 for mid-frequency and 125 for low-frequency (provided there were enough good candidates available, and with numbers varying a little as dictionary-derived candidates were added in later).

We found no dictionaries containing significant numbers of collocations for Czech. For English we used OCD, BBI (Benson et al., 2010), Macmillan Collocations Dictionary (Rundell, 2010), Oxford Dictionary of English,[5] Collins English Dictionary,[6] Wordnet,[7] and Merriam Webster.[8] Each headword in the sample was checked, with all collocations found in its entry added to the set of candidates.

The corpora and processing tools were as shown in Table 2. The three TenTen corpora are recent, web-crawled corpora created using similar methods to ukWaC (Ferraresi and Zanchetta, 2008). The SYN family corpora were all created and provided for this exercise by the Czech National Corpus project (ICNC, 2000 2013) and were processed by Morče (Hajič et al., 2007). Czes2 comprises newspaper and magazines, and we evaluated in three versions, all processed with the Desamb tagger (Šmerk, 2004; Šmerk, 2008), but two then further processed by parsers Synt (Jakubíček et al., 2009) and SET (Kovář et al., 2011). CzechParl (Jakubíček and Kovář, 2010) is a corpus of stenographic protocols from Czech parliament and was processed with the Desamb tagger too.

Corpus	Size	Tools
Czech		
SYN	2,232	Morče
SYN2010	121	Morče
SYN2009PUB	844	Morče
SYN2006PUB	361	Morče
SYN2005	122	Morče
SYN2000	120	Morče
CzechParl	45	Desamb
Czes2	368	Desamb
Czes2-SET	368	Desamb+Set
Czes2-Synt	368	Desamb+Synt
czTenTen12	4,791	Desamb
English		
BNC	96	CLAWS
ukWaC	1319	TreeTagger
enTenTen08	2,759	TreeTagger
enTenTen12	11,192	TreeTagger
NMCorpus	95	TreeTagger

Table 2: Corpora used for candidate generation, sizes in millions of words.

The other English corpora were the British National Corpus[9] and the NM Corpus, which is designed as an update of the BNC.

The unparsed Czes2 corpus, all the SYN corpora and all the English corpora used regular expressions over part-of-speech tags for the grammatical component of identifying collocations. Czes2-SET and Czes2-Synt used the SET and Synt parsers to produce a labeled dependency tree (SET) and an unlabeled dependency graph (Synt) and these were used for the grammatical component.

For English, by CLAWS we mean the CLAWS tokeniser and POS-tagger as in the published edition of the BNC[10] and the grammar described in (Kilgarriff et al., 2004); by TreeTagger, TreeTagger (Schmid, 1994) with the default model for English.[11]

For each corpus and each headword, we generated a stage-1 list of all collocations which occurred five or more times and which had a Dice coefficient indicating a positive association between the lemmas.

We then generated two stage-2 lists, one using raw frequency of collocation to order candidates, the other using the Dice coefficient. We then generated the stage-3 list (of length 500, 250 or 125, depending on the frequency-band of the headword) by taking one collocation from each stage-2 list in turn, adding it to the candidate list if it was not already there, otherwise moving on, until we had the target number.

[4] Adverbs are a far smaller category, usually accounting for less than 1% of dictionary headword lists.

[5] Checked at `oxforddictionaries.com`

[6] Checked at `www.collinsdictionary.com`

[7] Checked at `wordnetweb.princeton.edu/perl/webwn`

[8] Checked at `www.merriam-webster.com`

[9] See `natcorp.ox.ac.uk`

[10] See `natcorp.ox.ac.uk`

[11] As downloaded from `www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`

Rank	**hi** (100–2999)	**mid** (3000–9999)	**low** (10,000–30,000)
nouns	building circuit classroom close description distribution meeting metal participant percentage prayer rail virus vision wedding	bolt broadcast calorie editorial flame gauge maximum onset poisoning ram sediment showing telescope weed	blunder commoner democrat fitter hack harp mint saturation saying scuba semantics sewing slaughterhouse topography trawler
verbs	associate climb identify lecture like love matter top value view	contest empty inject instruct pile root rush slow tire	bathe dupe excrete glue instigate kid limp manoeuvre overshadow shelter
adjs	average black clean critical cultural disabled free global operational past	comic delicate intriguing lightweight loyal semantic stimulating supportive worthwhile	attainable delirious evocative pointed popup sublime tempting uncanny unofficial virulent

Table 1: The sampling frame, and English sample.

The final candidate list, to be shown to the judges, was then the merge of the stage-3 list and the dictionary list, randomised.[12]

5.3. Judging

For Czech, the judges were three Czech native speakers and students of linguistics, experienced in linguistic annotation. For English, the judges were three English native speakers and professional lexicographers, all of whom had worked on the 2nd edition of the Oxford Collocations Dictionary. In preliminary, standardisation exercises for each language, several words were judged by the judges and the native-speaker co-authors and discrepancies were discussed, so, as far as possible, we all agreed what was to count as good. The judging was undertaken at a web interface. Each headword had a separate page, with one row for each collocation. Collocation order was randomised. The row comprised the collocation, its commonest form (see above) and a choice of two boxes to tick: good or bad.[13] All judges assessed all collocations: 29,774 for Czech, 29,294 for English.

For Czech, the judges found 9.1%, 21.6% and 24.3% of collocations to be good. The agreement level between pairs of judges varied between 73.6% and 82.3%. Kappa ranged between 0.09 and 0.50.

For English, the judges found 15.8% 18.3% and 26.3% of collocations to be good. The agreement level between pairs of judges varied between 81.1% and 85.8%, with kappa between 0.44 and 0.50.

We treated all collocations which all but one of the judges had called 'good', as good.[14] A collocation got into the gold standard if it had, "three goods or two goods and a bad".

One way to investigate our success in finding all the collocations is to see if we would have found more, had we made the candidate lists longer. To examine this, we:

- ordered collocations according to their scores in the step-2 lists
- divided the list for each headword into fiftieths
- examined how many of the good collocates came from each fiftieth.[15]

Figure 1 shows that there were diminishing returns from asking the judges to judge more candidates identified with the same method and sources. Most of the good collocations were from the top fiftieths, with few from the lowest.

5.4. The Gold-Standard Datasets

For Czech the gold standard collocation set comprises 4,854 collocations for 100 headwords, and for English, the set comprises 5,327 collocations for 102 headwords. For Czech there were 5 headwords for which no collocations were found; for English there were none.

The distribution associated with this paper comprises a README describing data formats and, for each language:

- one file with the gold-standard collocations: ⟨headword, collocation⟩ pairs,
- one file with the full set of judgements: n-tuples ⟨headword, wordclass, freqband, collocate, list of judgements, rank⟩ and
- the stoplist.

6. The Corpus Evaluation

For both Czech and English, we evaluated the corpora used to generate the candidate collocations (see Table 2) plus, for English, the BNC processed with TreeTagger, the Oxford English Corpus (OEC),[16] and the 83 billion word enClueWeb09 corpus compiled from the English part of the ClueWeb09 collection.[17] For Czech, we added 2 more corpora processed by parsers: Czes2 processed by Morče and MaltParser (Nivre et al., 2006), and the same corpus processed by MST parser (McDonald et al., 2005), both trained on Prague Dependency Treebank (Hajič, 2006). We revisit the validity of comparing corpora that were, and were not,

[12] We found a common form for dictionary-only-sourced items so judges would not be aware which items were found in a dictionary, which in a corpus.

[13] There was also a 'show concordance' button, which, in practise, was almost never used, as the lookup took too long and the critical information was already available in the commonest form.

[14] We also explored only counting collocations which all judges liked, as good. This gave a smaller gold standard set and less stable results. To keep the number of parameters in check, we do not present results for this setting.

[15] Collocations that came only from the dictionary were set aside, for this exercise.

[16] www.oxforddictionaries.com/words/about-the-oxford-english-corpus

[17] lemurproject.org/clueweb09

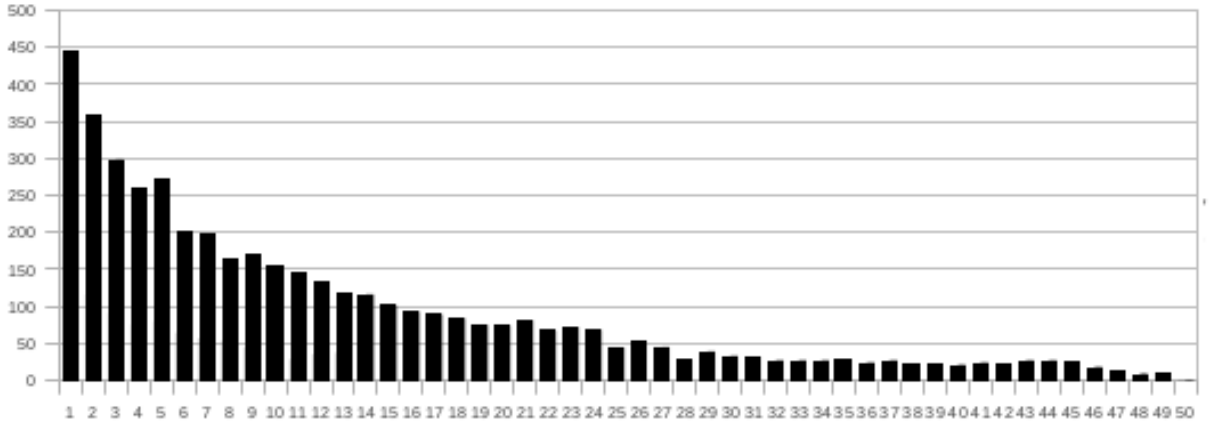

Figure 1: Distribution of good collocations in fiftieths, ordered by score.

	Result set size			
frq band	hi	mid	lo	hl
Hi	400	200	100	25
Mid	200	100	50	12
Lo	100	50	25	6

Table 3: Result set sizes, by frequency band

used to create the candidate set, below. For each corpus, we experimented with the following settings.

- Collocation sorting: by frequency, (fr) or by Dice coefficient (di).
- How big a result set to examine, for each headword. Clearly, the smaller sets will gain better precision, the bigger sets better recall. We have introduced static constraints, e.g. first 20 or 100 collocates, as well as variable thresholds dependent on the frequency band of the headword, with four possible values, *hi, mid, lo, hl*, as in Table 3. "-" in the result tables means no restriction, all collocates taken into account.
- the minimum number of hits for a collocation: 3, 5 or 10.

To indicate the parameters for a run, we run these together, e. g., fr/lo/5.

In lexicography recall is a greater challenge than precision. It is not so hard to check data and filter out unwanted items: finding all the instances of interest is much more important. Thus, for the evaluation, we wanted to evaluate both precision and recall, but to give greater weight to recall. To this end we have given scores according to F-5.[18] In tables 4 and 5 we present 3 best scores according to precision, recall and F-5 score. Table 6 contains the best F-5 scores for all English corpora from the testing set.

7. Discussion

With appropriate settings, we were able to achieve about 90 percent recall (but with very low precisions and high numbers of extracted collocations) and nearly 60 percent precision (40 for Czech). However, the best precision results achieve only 11–14 percent recall, so we do not consider them practically usable. The F-5 measure provides a good compromise between these two.

For precision, corpora prepared by careful selection of sources score better, such as the 100 million word BNC or NMCorpus.[19] For Czech, the parsed corpora achieved best precision, but the SYN corpus (part of Czech National Corpus, 2.2 billion words) scored well on recall and F-5 measure.

Size matters. Big web corpora are the general winners according to the F-5 score — namely the 83 billion word enClueWeb09, and the enTenTen pair of corpora (3 and 13 billion words), as well as the Czech member of the family, 5 billion word czTenTen. The 2.2 billion word SYN corpus scored also well.

Another important (and surprising) observation is that raw frequency is better than the Dice score for collocation salience estimations — it outperformed the Dice score practically in all important measures.

For Czech, parsing slightly helps with precision but has negative impact on recall. According to both recall and F-5, Czes2 processed by the parsers scored worse than the same corpus with the regular expressions grammar. There are only slight differences among the used parsers, as illustrated by Table 7. It is interesting that the parser scores do not correlate with results of usual parser evaluations, using tree similarities against the Prague Dependency Treebank.

7.1. Corpora Not Used for Candidate Generation, and Just-in-Time Evaluation

As OEC was a large, recent, high-quality corpus with a high level of investment, it was initially surprising to see its rather low score. It seemed likely that this was because it had not been used, as a source for generating the collocation candidates.

We explored the hypothesis as follows. For fifty of the English headwords, we identified the twenty highest-scoring collocations found in OEC but which had not occurred with high frequency or salience in the other corpora, so had not

[18] $F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$, where $\beta = 5$.

[19] www.sketchengine.co.uk/documentation/wiki/Corpora/NewModelCorpus

Corpus	Settings	# collocs	Prec (%)	Rec (%)	F-5 (%)
BNC	fr/hl/10	1,122	58.5	12.3	12.7
Model	fr/20/10	969	55.7	10.1	10.5
enTenTen12	fr/hl/10	1,456	52.7	14.4	14.8
enClueWeb09	fr/-/3	89,174	5.4	90.1	56.6
enTenTen12	fr/-/3	66,499	7.2	89.5	62.1
enTenTen08	fr/-/3	52,018	8.9	87.1	65.1
enTenTen12	fr/hi/5	23,113	17.8	77.3	68.5
enTenTen08	fr/hi/3	22,794	18.0	77.1	68.5
enClueWeb09	fr/hi/10	23,651	17.4	77.0	68.0

Table 4: Best results of English evaluation: 3 best according to precision, 3 best according to recall and 3 according to F-5.

Corpus	Settings	# collocs	Prec (%)	Rec (%)	F-5 (%)
Czes2 MST	di/hl/10	1,480	38.8	11.8	12.2
Czes2 Malt	di/hl/10	1,478	38.0	11.6	11.9
Czes2	di/hl/10	1,447	36.8	11.0	11.3
czTenTen12	fr/-/3	58,462	7.2	86.7	60.9
SYN	fr/-/3	52,000	7.9	85.1	62.0
Czes2	fr/-/3	32,419	11.0	73.4	60.3
czTenTen12	fr/-/10	40,654	9.9	82.9	64.6
SYN	fr/-/10	34,667	11.1	79.1	64.0
Czes2	fr/-/3	32,419	11.0	73.4	60.3

Table 5: Best results of Czech evaluation: 3 best according to precision, 3 best according to recall and 3 according to F-5.

been in the original candidate set. We then asked the same three judges to judge these items.

Of 984 collocations judged, 187 (19%) were judged good by at least two of the three judges. This 'overall good' rate is close to the rate for the original candidate set. The hypothesis that OEC scored badly because it was not used as a source for the original candidate list is confirmed. As it stands, the gold standard dataset only serves to evaluate those corpora that have been used to build the candidate set. As explored above, adding to the candidate set by showing the judges more candidates from existing corpora will find few additional collocations. However showing them additional candidates from other corpora will find many. This is in keeping with a common finding in corpora: the more you look, the more you find.[20]

It suggests an extension to the framework to support evaluation of additional corpora: a 'just-in-time' method where we identify those candidates that would have been in the collocation set, had the corpus been included originally, and show them to judges. Then we can use the extended gold-standard to compare the new corpus with the original set.

Had we included OEC amongst the original corpora (and allowed the candidate set sizes to extend beyond 500, 250, 125), then we would have replaced 3254 items in the candidate set. For the modest cost of getting additional judgement on (in this case) 3254 candidates, we can include an extra corpus in the set to be compared.

[20]There is one other explanation to be explored: that judges make their judgements relative to the candidate set they are shown, so will use tighter criteria if shown a better candidate set and more relaxed ones if shown a worse one. We shall be covering this possibility in future work.

8. Conclusion and Further Work

Corpus evaluation is critical to the progress of the field. Cleverer and cleverer programs operating on the same old flawed data will not get us far, but we will only know it is flawed if we can evaluate it. We have presented an approach to evaluating general-language corpora based around the question "how good is this corpus for creating a publication-quality collocations dictionary?". For 100 headwords of Czech, and 100 of English, expert judges identified (as far as possible) all the collocations that should go into such a dictionary, and this gold standard set, which has been included with the paper, was then used to evaluate a set of corpora for each language.

Our original, most optimistic hope was that we might gather a complete set of 'good' collocations for the headwords. This turned out to be unrealistic because

- if we showed judges more candidates from the same corpora, they found more collocations (though with diminishing returns)
- if we showed judges more candidates from new corpora, they found more collocations.

This weakens our claims to establish recall, but still allows us to compare corpora. To compare an additional corpus to the ones used to prepare the candidate set, we send extra collocations to the judges for just-in-time evaluation.

Now that we have the framework, we (and, we hope, others) shall use it in a number of ways:

- to set parameters for data cleaning and deduplication, in our corpus-building
- to evaluate different crawling strategies

Corpus	Settings	# collocs	Prec (%)	Rec (%)	F-5 (%)
enTenTen12	fr/hi/5	23,113	17.8	77.3	68.5
enTenTen08	fr/hi/3	22,794	18.0	77.1	68.5
enClueWeb09	fr/hi/10	23,651	17.4	77.0	68.0
ukWaC	fr/hi/3	22,287	17.6	73.7	65.7
OEC	fr/hi/3	22,832	17.0	72.7	64.5
BNC (CLAWS)	fr/-/3	13,158	20.7	51.2	48.5
NMCorpus	fr/-/3	13,170	20.4	50.4	47.7
BNC (TreeTagger)	fr/-/3	12,944	20.5	49.7	47.1

Table 6: Best results for each of the English corpora, according to F-5.

Corpus	Settings	# collocs	PDT score (%)	Prec (%)	Rec (%)	F-5 (%)
Czes2	fr/-/3	32,419	N/A	11.0	73.4	60.3
Czes2 SET	fr/-/3	35,729	56.0	9.7	71.2	57.2
Czes2 Synt	fr/-/3	18,708	N/A	15.5	59.9	54.0
Czes2 MST	fr/-/3	32,581	84.7	10.5	70.5	57.8
Czes2 Malt	fr/-/3	32,471	85.8	10.5	70.2	57.6

Table 7: Comparison of Czech parsers using collocation extraction, and using dependency precision against Prague Dependency Treebank (PDT). Settings selected according to best F-5 score.

- to compare different processing chains
- to evaluate grammars.

9. Acknowledgements

This work has been partly supported by the Ministry of Education of the Czech Republic within the LINDAT-Clarin project LM2010013 and by the Ministry of the Interior of the Czech Republic within the project VF20102014003.

10. References

Benson, M., Benson, E., and Ilson, R. (2010). *The BBI Combinatory Dictionary of English, 3rd edition*. John Benjamins.

Crowther, J., Dignen, S., and Lea, D. (2002). *Oxford Collocations Dictionary for Students of English*. Oxford University Press.

Ferraresi, A. and Zanchetta, E. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA)*.

Gaizauskas, R. (1998). Evaluation in language and speech technology. *Journal of Computer Speech and Language*, 12(3):249–262.

Hajič, J., Votrubec, J., Krbec, P., Květoň, P., et al. (2007). The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 67–74. Association for Computational Linguistics.

Hajič, J. (2006). Complex corpus annotation: The Prague dependency treebank. *Insight into the Slovak and Czech Corpus Linguistics*, page 54.

ICNC. (2000–2013). *Czech National Corpora – SYN, SYN2000, SYN2005, SYN2006PUB, SYN2009PUB, SYN2010*. Institute of Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic. Available online at `www.korpus.cz`.

Jakubíček, M., Kovář, V., and Horák, A. (2009). Mining phrases from syntactic analysis. In *Lecture Notes in Artificial Intelligence, Proceedings of Text, Speech and Dialogue 2009*, pages 124–130, Plzeň, Czech Republic. Springer-Verlag.

Jakubíček, M. and Kovář, V. (2010). CzechParl: Corpus of stenographic protocols from Czech parliament. *RASLAN 2010 Recent Advances in Slavonic Natural Language Processing*, page 41.

Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The Sketch Engine. *Information Technology*, 105:116.

Kilgarriff, A., Kovář, V., Krek, S., Srdanović, I., and Tiberius, C. (2010). A quantitative evaluation of word sketches. In *Proceedings of the XIV Euralex International Congress, Leeuwarden: Fryske Academy*.

Kilgarriff, A., Rychlý, P., Kovář, V., and Baisa, V. (2012). Finding multiwords of more than two words. In *Proceedings of the 15th EURALEX International Congress*, pages 693–700.

Kovář, V., Horák, A., and Jakubíček, M. (2011). Syntactic analysis using finite patterns: a new parsing system for Czech. *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 161–171.

Krenn, B., Evert, S., et al. (2001). Can we do better than frequency? A case study on extracting pp-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46.

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, BC, Canada. Association for Computational Linguistics.

Nivre, J., Hall, J., and Nilsson, J. (2006). MaltParser:

A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC 2006)*, volume 6, Genoa, Italy. European Language Resource Association, Paris.

Rundell, M. (2010). *Macmillan Collocations Dictionary*. Macmillan.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.

Šmerk, P. (2004). Unsupervised Learning of Rules for Morphological Disambiguation. In *Lecture Notes in Artificial Intelligence 3206, Proceedings of Text, Speech and Dialogue 2004*, pages 211–216, Berlin. Springer-Verlag.

Šmerk, P. (2008). Towards Czech morphological guesser. *Sojka, Petr-Horák, Aleš. Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pages 1–4.