Extrinsic corpus evaluation with a collocation dictionary task

Adam Kilgarriff
Lexical Computing, Ltd.
adam@lexmasterclass.com

Pavel Rychlý
Masaryk University
pary@fi.muni.cz

Miloš Jakubíček
Masaryk University
xjakub@fi.muni.cz

Vojtěch Kovář
Masaryk University
xkovar3@fi.muni.cz

Vít Baisa
Masaryk University
xbaisa@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.

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.

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

[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 http://trec.nist.gov, in particular http://trec.nist.gov/presentations/TREC5/15.html

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 (ANON)); whatever the collocation, we show the judge *male chauvinist pig* together wiht 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.

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 ajectives 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).

[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.

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

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.

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.

Corpus	**Size**	**Tools**
Czech		
SYN	1,568	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 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.

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.

[5] Checked at http://oxforddictionaries.com/

[6] Checked at http://www.collinsdictionary.com/

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

[8] Checked at http://www.merriam-webster.com/

[9] See http://natcorp.ox.ac.uk

[10] See http://natcorp.ox.ac.uk

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

We then generated two stage-2 lists, one using raw frequency of collocation to order candidates, the other using the dice co-efficient. 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.

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 four Czech native speakers and students of linguistics. 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 four judges found 3.8% 9.1%, 21.6% and 24.3% of collocations to be good. The agreement level between pairs of judges varied between 73.6% and 89.7% (although this last score is between the two judges who found very few collocations to be good, so the high agreement score disguises a low Cohen's kappa, of 0.15). Kappa ranged between 0.09 and 0.50.

For English, the four 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.

[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.

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, for Czech, "four goods or three goods and a bad" and for English, "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 1378 collocations for 85 headwords, and for English, the set comprises 5,327 collocations for 102 headwords. For Czech there were 20 headowrds for whcih no collocations were found; for English there were none. The highest, median, and lowest number of collocations per headword, by frequency band and word class, is shown in Table 3.

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.

[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.

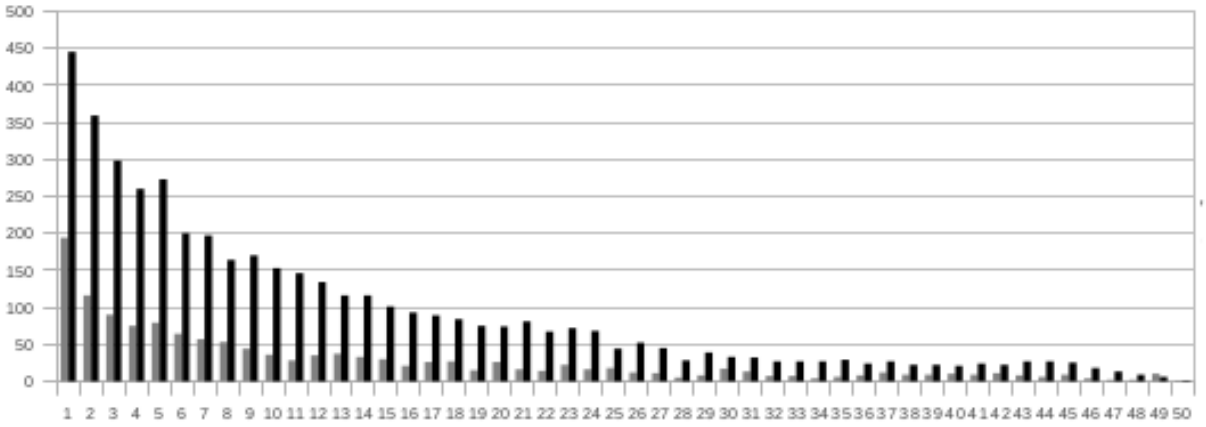

Figure 1: Distribution of good collocations in fiftieths, ordered by score. English is black, Czech grey.

		English						Czech					
		hi		mid		lo		hi		mid		lo	
		word	#	word	#	word	#	word	#	word	#	word	#
n	max	building	199	blunder	63	flame	85	důkaz	103	box	41	hadička	19
	med	classroom	90	topography	18	gauge	38	federace	18	nájezd	15	ilustrátor	4
	min	participant	36	commoner	4	ram	21	příslušník	3	zaplacení	1	metrák	1
j	max	average	176	delicate	67	evocative	43	dopravní	61	dokončený	18	huňatý	24
	med	black	118	worthwhile	61	tempting	25	minimální	23	pedagogický	4	ušitý	2
	min	operational	49	semantic	24	popup	12	složitý	1	časný	1	posedlý	2
v	max	identify	95	instigate	58	attribute	91	jednat	19	dýchat	37	vyhazovat	12
	med	matter	45	shelter	15	inject	30	požádat	9	naplánovat	9	zaleknout	1
	min	like	20	kid	8	tire	7	způsobit	2	zkrátit	1	odstát	1

Table 3: For each language, word class and frequency band. the highest, median and lowest number of collocations for a heaword, and the associated headword.

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, and the Oxford English Corpus. We revisit the validity of comparing corpora that were, and were not, used to create the candidate set, below. For each corpus, we experimented with the following settings.

- collocation sorting: by frequency, (fr) or by dice co-efficient (di)
- how big a result set to examine, for each headword. This was a variable threshold dependent on the frequency band of the headword, with three possible values, *Hi, Mid, Lo*, as in Table 4
- the minumum number of hits for a collocation: 1, 5 or 10.

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

	Result set		
frq band	Hi	Mid	Lo
Hi	400	200	100
Mid	200	100	50
Lo	100	50	25

Table 4: Result set sizes, by frequency band

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 a much harder task. 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[16]. In tables 5 and 6 we present each corpus, the parameters for it that scored highest according to F-5, and the F-5 score.

7 Discussion

Size matters. For English, the three highest scorers (excluding OEC, see below) were the three largest

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

corpus	params	prec	rec	F-5
Czes2-Synt	di/lo/5	11.60	47.46	42.42
Czes2-SET	di/lo/5	10.40	44.05	39.18
SYN	di/lo/10	9.05	38.46	34.19
czTenTen12	fr/lo/10	8.71	37.88	33.56
SYN2009PUB	di/lo/10	9.47	37.30	33.51
SYN2006PUB	di/lo/5	9.13	36.65	32.84
SYN2010	di/lo/5	10.88	35.63	32.76
Czes2	fr/lo/10	10.13	35.78	32.60
SYN2005	di/lo/5	10.64	35.41	32.50
SYN2000	di/mid/10	11.72	28.81	27.28
CzechParl	di/lo/5	9.88	14.95	14.66

Table 5: Evaluation of Czech corpora.

corpus	params	prec	rec	F-5
enTenTen12	fr/lo/1	30.95	34.43	34.28
enTenTen08	fr/lo/5	30.94	34.18	34.05
ukWAC	fr/lo/5	29.83	32.76	32.63
BNC (tt)	fr/lo/1	26.51	29.32	29.20
BNC	fr/lo/1	26.21	28.97	28.85
NMCorpus	fr/lo/1	25.67	28.55	28.43
OEC	fr/lo/10	28.65	28.12	28.14
ACL ARC	fr/lo/1	14.21	11.90	11.98

Table 6: Evaluation of English corpora.

corpora, in size order. For Czech, once parsed corpora are set aside, a similar if weaker relation holds.

For Czech, parsing helps. Both versions of Czes2 that used a parser substantially outperformed the version that did not use a parser, and all other Czech corpora. This is decisive evidence for the benefits of parsing.

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 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 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.[17]

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.

For the Czech corpora, both parsers clearly outperform the CQL grammar based on regular expressions, which is not as evident as it might seem at the first sight: similar attempts gained ambivalent results as far (Horák et al., 2009; Ambati et al., 2012), thus not being convincing that the (significantly larger) processing time and related interoperability issues pay off.

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.

[17] 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.

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
- to compare different processing chains
- to evaluate grammars.

References

Bharat Ram Ambati, Siva Reddy, and Adam Kilgarriff. 2012. Word sketches for Turkish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, May. European Language Resources Association (ELRA)*.

Morton Benson, Evelyn Benson, and Robert Ilson. 2010. *The BBI Combinatory Dictionary of English, 3rd edition*. John Benjamins.

Jonathan Crowther, Sheila Dignen, and Diana Lea. 2002. *Oxford Collocations Dictionary for Students of English*. Oxford University Press.

Adriano Ferraresi and Eros Zanchetta. 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)*.

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

Jan Hajič, Jan Votrubec, Pavel Krbec, Pavel Květoň, 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.

A. Horák, P. Rychlý, and A. Kilgarriff. 2009. Czech word sketch relations with full syntax parser. *After Half a Century of Slavonic Natural Language Processing*, pages 101–112.

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

M. Jakubíček, V. Kovář, and A. Horák. 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.

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

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. Itri-04-08 the sketch engine. *Information Technology*, 105:116.

Adam Kilgarriff, Vojtech Kovar, Simon Krek, Irena Srdanovic, and Carole Tiberius. 2010. A quantitative evaluation of word sketches. In *Proceedings of the XIV Euralex International Congress, Leeuwarden: Fryske Academy*.

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

Brigitte Krenn, Stefan Evert, 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.

Michael Rundell. 2010. *Macmillan Collocations Dictionary*. Macmillan.

Helmut Schmid. 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.

Pavel Šmerk. 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.

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