{bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian

Nikola Ljubešić
University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
nljubesi@ffzg.hr

Filip Klubička
University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
fklubick@ffzg.hr

Abstract

In this paper we present the construction process of top-level-domain web corpora of Bosnian, Croatian and Serbian. For constructing the corpora we use the SpiderLing crawler with its associated tools adapted for simultaneous crawling and processing of text written in two scripts, Latin and Cyrillic. In addition to the modified collection process we focus on two sources of noise in the resulting corpora: 1. they contain documents written in the other, closely related languages that can not be identified with standard language identification methods and 2. as most web corpora, they partially contain low-quality data not suitable for the specific research and application objectives. We approach both problems by using language modeling on the crawled data only, omitting the need for manually validated language samples for training. On the task of discriminating between closely related languages we outperform the state-of-the-art Blacklist classifier reducing its error to a fourth.

1 Introduction

Building web corpora for various NLP tasks has become quite a standard approach, especially if funding is limited and / or there is need for large amounts of textual data.

Although off-the-shelf solutions for compiling web corpora have emerged recently, there are still specific challenges that have to be addressed in most corpus construction processes. One such challenge that we face while constructing the corpora described in this paper is simultaneous usage of two scripts on two out of three top-level domains (TLDs) crawled.

Additionally, there are still many open questions and possibilities for improvement in the process of collecting data as well as data post-processing. We address two of the latter kind – discrimination between similar, neighboring languages that are used on all selected TLDs, and the question of text quality in corpora collected in such a fully automated fashion.

In the paper we present the process of building web corpora of Bosnian, Croatian and Serbian by crawling the .ba, .hr and .rs TLDs. The three languages belong to the South Slavic language branch and are very similar to each other. The biggest differences between Croatian and Serbian are the proto-Slavic vowel *jat* (Croatian *čovjek* vs. Serbian *čovek*), way of handling proper nouns (Croatian *New York* vs. Serbian *Nju Jork*), specific syntactic constructions (Croatian *hoću raditi* vs. Serbian *hoću da radim*) and a series of lexical differences (Croatian *mrkva* vs. Serbian *šargarepa*). Bosnian is mostly seen as a mixture of those two and allows, beside its own lexical specificities, solutions from one or both languages.[1]

This paper is structured as follows: in Section 2 we give an overview of related work regarding existing (web) corpora of the languages in question, language identification and web text quality estimation. Section 3 shows the process of collecting the three TLD corpora with emphasis on the problem of collecting data written in various scripts, while in Section 4 we describe the linguistic annotation layers added to the corpora. Section 5 depicts our approach to discriminating between very similar languages while in Section 6 we describe our approach to identifying documents of low text quality, and both approaches use recently crawled data only.

[1] A more thorough comparison of the three languages is available at http://en.wikipedia.org/wiki/Comparison_of_standard_Bosnian,_Croatian_and_Serbian

Felix Bildhauer & Roland Schäfer (eds.), *Proceedings of the 9th Web as Corpus Workshop (WaC-9) @ EACL 2014*, pages 29–35,
Gothenburg, Sweden, April 26 2014. ©2014 Association for Computational Linguistics

2 Related work

The only two South Slavic languages for which web corpora were previously built are Croatian and Slovene (Ljubešić and Erjavec, 2011). The Croatian corpus presented in this paper is actually an extension of the existing corpus, representing its second version. hrWaC v1.0 was, until now, the biggest available corpus of Croatian.

For Bosnian, almost no corpora are available except the SETimes corpus[2], which is a 10-languages parallel corpus with its Bosnian side consisting of 2.2 million words, and The Oslo Corpus of Bosnian Texts[3], which is a 1.5 million words corpus consisting of different genres of texts that were published in the 1990s.

For the Serbian language, until now, the largest corpus was the SrpKor corpus[4], consisting of 118 million words that are annotated with part-of-speech information (16 tags) and lemmatized. The corpus is available for search through an interface for non-commercial purposes.

Until now, no large freely downloadable corpora of Bosnian and Serbian were available, and this was one of the strongest motivations for our work.

Multiple pipelines for building web corpora were described in many papers in the last decade (Baroni et al., 2009; Ljubešić and Erjavec, 2011; Schäfer and Bildhauer, 2012), but, to the best of our knowledge, only one pipeline is freely available as a complete, ready-to-use tool: the Brno pipeline (Suchomel and Pomikálek, 2012), consisting of the SpiderLing crawler[5], the Chared encoding detector[6], the jusText content extractor[7] and the Onion near-deduplicator[8]. Although we have our own pipeline set up (this is the pipeline the first versions of hrWaC and slWaC were built with), we decided to compile these versions of web corpora with the Brno pipeline for two reasons: 1. to inspect the pipeline's capabilities, and 2. to extend the Croatian web corpus as much as possible by using a different crawler.

Although language identification is seen as a solved problem by many, the recently growing interest for it indicates the opposite. Recently, researchers focused on improving off-the-shelf tools for identifying many languages (Lui and Baldwin, 2012), discriminating between similar languages where standard tools fail (Tiedemann and Ljubešić, 2012), identifying documents written in multiple languages and identifying the languages in such multilingual documents (Lui et al., 2014).

Text quality in automatically constructed web corpora is quite an underresearched topic, with the exception of boilerplate removal / content extraction approaches that deal with this problem implicitly (Baroni et al., 2008; Kohlschütter et al., 2010), but quite drastically, by removing all content that does not conform to the criteria set. A recent approach to assessing text quality in web corpora in an unsupervised manner (Schäfer et al., 2013) calculates the weighted mean and standard deviation of n most frequent words in a corpus sample and measures how much a specific document deviates from the estimated means. This approach is in its basic idea quite similar to ours because it assumes that most of the documents in the corpus contain content of good quality. The main difference in our approach is that we do not constrain ourselves to most frequent words as features, but use character and word n-grams of all available text.

3 Corpus construction

For constructing the corpora we used the SpiderLing crawler[9] along with its associated tools for encoding guessing, content extraction, language identification and near-duplicate removal (Suchomel and Pomikálek, 2012). Seed URLs for Bosnian and Serbian were obtained via the Google Search API queried with bigrams of mid-frequency terms. Those terms were obtained from corpora that were built with focused crawls of newspaper sites as part of our previous research (Tiedemann and Ljubešić, 2012). For Croatian seed URLs, we used the home pages of web domains obtained during the construction of the first version of the hrWaC corpus. The number of seed URLs was 8,388 for bsWaC, 11,427 for srWaC and 14,396 for hrWaC. Each TLD was crawled for 21 days with 16 cores used for document processing.

Because Serbian – which is frequently used on the Serbian and Bosnian TLDs – uses two scripts

[2] http://nlp.ffzg.hr/resources/corpora/setimes/

[3] http://www.tekstlab.uio.no/Bosnian/Corpus.html

[4] http://tinyurl.com/mocnzna

[5] http://nlp.fi.muni.cz/trac/spiderling

[6] https://code.google.com/p/chared/

[7] http://code.google.com/p/justext/

[8] http://code.google.com/p/onion/

[9] http://nlp.fi.muni.cz/trac/spiderling

– Latin and Cyrillic – we had to adjust the standard corpus construction process to cope with both scripts. This was done by 1. building new two-script models for encoding guessing with Chared, 2. defining stop-words used in content extraction in both scripts and 3. transforming extracted text from Cyrillic to Latin with *serbian.py*[10] before performing language identification and duplicate removal. We kept all content of the final corpora in the Latin script to simplify further processing, especially because linguistic annotation layers were added with models developed for Croatian which uses the Latin script exclusively. The information about the amount of Cyrillic text in each document is still preserved as an attribute of the `<doc>` element. Overall the percentage of documents written >90% in the Cyrillic script was 3.2% on the Bosnian TLD and 16.7% on the Serbian TLD.

Near-duplicate identification was performed both on the document and the paragraph level. The document-level near-duplicates were removed from the corpus cutting its size in half, while paragraph-level near-duplicates were labeled by the `neardupe` binary attribute in the `<p>` element enabling the corpus users to decide what level of near-duplicate removal suits their needs.

The resulting size of the three corpora (in millions of tokens) after each of the three duplicate removal stages is given in Table 1. Separate numbers are shown for the new crawl of the Croatian TLD and the final corpus consisting of both crawls.

	PHYS	DOCN	PARN
bsWaC 1.0	722	429	288
hrWaC new	1,779	1,134	700
hrWaC 2.0	2,686	1,910	1,340
srWaC 1.0	1,554	894	557

Table 1: Size of the corpora in Mtokens after physical duplicate (PHY), document near-duplicate (DOCN) and paragraph near-duplicate removal (PARN)

At this point of the corpus construction process the `<doc>` element contained the following attributes:

- `domain` – the domain the document is published on (e.g. `zkvh.org.rs`)
- `url` – the URL of the document
- `crawl_date` – date the document was crawled
- `cyrillic_num` – number of Cyrillic letters in the document
- `cyrillic_perc` – percentage of letters that are Cyrillic

4 Corpus annotation

We annotated all three corpora on the level of lemmas, morphosyntactic description (675 tags) and dependency syntax (15 tags). Lemmatization was performed with the CST's Lemmatiser[11] (Jongejan and Dalianis, 2009), morphosyntactic tagging with HunPos[12] (Halácsy et al., 2007) and dependency syntax with mate-tools[13] (Bohnet, 2010). All models were trained on the Croatian 90k-token annotated corpus SETimes.HR[14] (Agić and Ljubešić, 2014) that we recently expanded with 50k additional tokens from various newspaper domains (at this point we call it simply SETimes.HR+). Although the annotated training corpora are Croatian, previous research (Agić et al., 2013a; Agić et al., 2013b) has shown that on this level of tagging accuracy on in-domain test sets (lemma ≈96%, morphosyntactic description (MSD) ≈87%, labeled attachment score (LAS) ≈73%), annotating Serbian text with models trained on Croatian data produced performance loss of only up to 3% on all three levels of annotation, while on out-of-domain test sets (lemma ≈92%, MSD ≈81%, LAS ≈65%) there was no loss in accuracy.

We nevertheless performed an intervention in the SETimes.HR+ corpus before training the models used for annotating the Bosnian and the Serbian TLD corpora. Namely, on the morphosyntactic level the tagsets of Croatian and Serbian are identical, except for one subset of tags for the future tense which is present in Serbian and not present in Croatian. This is because Croatian uses the complex, analytic future tense consisting of the infinitive of the main verb and the present tense of the auxiliary verb *have* (*radit ćemo*) while Serbian uses both the analytic and the synthetic form where the two words are conflated into one (*radićemo*).

[10] `http://klaus.e175.net/code/serbian.py`

[11] `https://github.com/kuhumcst/cstlemma`

[12] `https://code.google.com/p/hunpos/`

[13] `https://code.google.com/p/mate-tools/`

[14] `http://nlp.ffzg.hr/resources/corpora/setimes-hr/`

To enable models to correctly handle both the analytic and synthetic form of the future tense, we simply repeated the sentences containing the analytic form that we automatically transformed to the synthetic one. By annotating the bsWaC and srWaC corpora with the models trained on the modified SETimes.HR+ corpus, we annotated 610k word forms in srWaC and 115k word forms in bsWaC with the synthetic future tense. Manual inspection showed that most of the tokens actually do represent the future tense, proving that the intervention was well worth it.

The lemmatization and morphosyntactic annotation of all three corpora took just a few hours while the full dependency parsing procedure on 40 server grade cores took 25 days.

5 Language identification

Because each of the three languages of interest is used to some extent on each of the three TLDs and, additionally, these languages are very similar, discriminating between them presented both a necessity and a challenge.

In previous work on discriminating between closely related languages, the Blacklist (BL) classifier (Tiedemann and Ljubešić, 2012) has shown to be, on a newspaper-based test set, 100% accurate in discriminating between Croatian and Serbian, and 97% accurate on all three languages of interest.

Our aim at this stage was twofold: 1. to put the existing BL classifier on a realistic test on (noisy) web data and 2. to propose an alternative, simple, data-intense, but noise-resistant method which can be used for discriminating between closely related languages or language varieties that are predominantly used on specific sections of the Web.

Our method (LM1) uses the whole content of each of the three TLD web corpora (so large amounts of automatically collected, noisy data) to build unigram-level language models. Its advantage over the BL classifier is that it does not require any clean, manually prepared samples for training. The probability estimate for each word w given the TLD, using add-one smoothing is this:

$$\hat{P}(w|TLD) = \frac{c(w, TLD) + 1}{\sum_{w_i \in V}(c(w_i, TLD) + 1)} \quad (1)$$

where $c(w, TLD)$ is the number of times word w occurred on the specific TLD and V is the vocabulary defined over all TLDs.

We perform classification on each document as a *maximum-a-posteriori* (MAP) decision, i.e. we choose the language of the corresponding TLD ($l \in TLD$) that produces maximum probability with respect to words occurring in the document ($w_1...w_n$):

$$l_{map} = \operatorname*{arg\,max}_{l \in TLD} \prod_{i=1..n} \hat{P}(w_i|l) \quad (2)$$

We should note here that our approach is identical to using the Naïve Bayes classifier without the *a priori* probability for each class, i.e. language.

Speaking in loose terms, what we do is that for each document of each TLD, we identify, on the word level, to which TLD data collection the document corresponds best.

Because Bosnian is mostly a mixture of Croatian and Serbian and actually represents a continuum between those two languages, we decided to compare the BL and the LM1 classifier on a much more straight-forward task of discriminating between Croatian and Serbian. The results of classifying each document with both classifiers are given in Table 2. They show that both classifiers agree on around 75% of decisions and that around 0.4 percent of documents from hrWaC are identified as Serbian and 1.5 percent of document from srWaC as Croatian.

	BL	LM1	agreement
hrWaC	0.42%	0.3%	73.15%
srWaC	1.93 %	1.28%	80.53%

Table 2: Percentage of documents identified by each classifier as belonging to the other language

We compared the classifiers by manually inspecting 100 random documents per corpus where the two classifiers were not in agreement. The results of this tool-oriented evaluation are presented in Table 3 showing that the LM1 classifier produced the correct answer in overall 4 times more cases than the BL classifier.

If we assume that the decisions where the two classifiers agree are correct (and manual inspection of data samples points in that direction) we can conclude that our simple, data-intense, noise-resistant LM1 method cuts the BL classification error to a fourth. We consider a more thorough evaluation of the two classifiers, probably by pooling and annotating documents that were identified

	BL	LM1	NA
hrWaC	18%	62%	20%
srWaC	10%	48%	42%

Table 3: Percentage of correct decisions of each classifier on documents where the classifiers disagreed (NA represents documents that are a mixture of both languages)

as belonging to the other TLD language by some classifier, as future work.

Due to the significant reduction in error by the LM1 classifier, we annotated each document in the hrWaC and srWaC corpora with the LM1 binary hr-sr language identifier while on bsWaC we used the LM1 ternary bs-hr-sr classifier. This decision is based on the fact that discriminating between all three languages is very hard even for humans and that for most users the hr-sr discrimination on the two corpora will be informative enough. In each document we encoded the normalized distribution of log-probabilities for the considered languages, enabling the corpus user to redefine his own language criterion.

The percentage of documents from each corpus being identified as a specific language is given in Table 4.

	bs	hr	sr
bsWaC	78.0%	16.5%	5.5%
hrWaC	-	99.7%	0.3%
srWaC	-	1.3%	98.7%

Table 4: Distribution of identified languages throughout the three corpora

Additional attributes added to the `<doc>` element during language identification are these:

- `lang` – language code of the language identified by maximum-a-posteriori
- `langdistr` – normalized distribution of log probabilities of languages taken under consideration (e.g. `bs:-0.324|hr:-0.329|sr:-0.347` for a document from bsWaC)

6 Identifying text of low quality

Finally, we tackled the problem of identifying documents of low text quality in an unsupervised manner by assuming that most of the content of each web corpus is of good quality and that low quality content can be identified as data points of lowest probability regarding language models built on the whole data collection. We pragmatically define low quality content as content not desirable for a significant number of research or application objectives.

For each TLD we calculated character n-gram and word n-gram language models in the same manner as in the previous section (Equation 1) for language identification. We scored each TLD document with each language model that was built on that TLD. To get a probability estimate which does not depend on the document length, we calculated probabilities of subsequences of identical length and computed the average of those.

We manually inspected documents with low probability regarding character n-gram models from level 1 to level 15 and word n-gram models from level 1 to level 5. Word n-gram models proved to be much less appropriate for capturing low quality documents by lowest probability scores than character n-gram models. Among character n-gram models, 3-gram models were able to identify documents with noise on the token level while 12-gram models assigned low probabilities to documents with noise above the token level.

The most frequent types of potential noise found in lowest scored documents in all three corpora are the following:

- 3-gram models
  - non-standard usage of uppercase, lowercase and punctuation
  - URL-s
  - uppercase want ads
  - formulas
- 12-gram models
  - words split into multiple words (due to soft hyphen usage or HTML tags inside words)
  - enumerated and bulleted lists
  - uppercase want ads
  - non-standard text (slang, no uppercased words, emoticons)
  - dialects
  - lyric, epic, historical texts

The character 3-gram method has additionally proven to be a very good estimate of text quality on the lexical level by strongly correlating (0.74) with the knowledge-heavy method of calculating lexical overlap of each document with a morphological dictionary which is available for Croatian[15].

An interesting finding is that word-level models perform much worse for this task than character-level models. We hypothesize that this is due to feature space sparsity on the word level which is much lower on the character level.

We decided to postpone any final decisions (like discretizing these two variables and defining one or two categorical ones) and therefore encoded both log-probabilities as attributes in each document element in the corpus leaving to the final users to define their own cut-off criteria. To make that decision easier, for each document and each character n-gram method we computed the percentage of documents in the corpus that have an equal or lower result of that character n-gram method. This makes removing a specific percentage of documents with lowest scores regarding a method much easier.

We also computed one very simple estimate of text quality – the percentage of characters that are diacritics. Namely, for some tasks, like lexicon enrichment, working on non-diacritized text is not an option. Additionally, it is to expect that lower usage of diacritics points to less standard language usage. The distribution of this text quality estimate in the hrWaC corpus (all three corpora follow the same pattern) is depicted in Figure 1 showing that the estimate is rather normally distributed with a small peak at value zero representing non-diacritized documents.

In each <doc> element we finally encoded 5 attributes regarding text quality:

- 3graph – average log-probability on 100-character sequences regarding the character 3-gram model trained on the whole TLD corpus

- 3graph_cumul – percentage of documents with equal or lower 3graph attribute value

- 12graph – same as 3graph, but computed with the character 12-gram model

- 12graph_cumul – like 3graph_cumul, but for the 12graph attribute

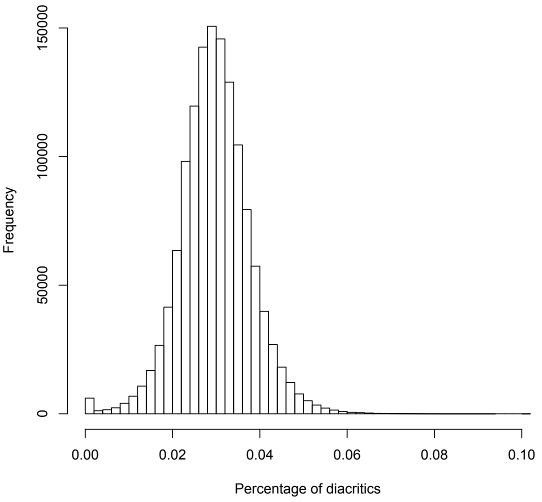

Figure 1: Distribution of the percentage of characters of a document being diacritics

- diacr_perc – percentage of non-whitespace characters that are diacritics

We plan to perform extrinsic evaluation of the three estimates of text quality on various NLP tasks such as language modeling for statistical machine translation, morphological lexicon induction, distributional lexicon induction of closely related languages and multi-word expression extraction.

7 Conclusion

In this paper we described the process of constructing three TLD corpora of Bosnian, Croatian and Serbian.

After presenting the construction and annotation process of the largest existing corpora for each of the three languages, we focused on the issue that all three languages are to some extent used on all three TLDs. We presented a method for discriminating between similar languages that is based on unigram language modeling on the crawled data only, which exploits the fact that the majority of the data published on each TLD is written in the language corresponding to that TLD. We reduced the error of a state-of-the-art classifier to a fourth on documents where the two classifiers disagree on.

We dealt with the problem of identifying low quality content as well, again using language modeling on crawled data only, showing that document probability regarding a character 3-gram model is a very good estimate of lexical quality, while low

[15] http://bit.ly/1mRjMrP

character 12-gram probabilities identify low quality documents beyond the word boundary.

We encoded a total of 12 attributes in the document element and the paragraph-near-duplicate information in the paragraph element enabling each user to search for and define his own criteria.

We plan on experimenting with those attributes on various tasks, from language modeling for statistical machine translation, to extracting various linguistic knowledge from those corpora.

Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement no. PIAP-GA-2012-324414 (project Abu-MaTran).

References

[Agić and Ljubešić2014] Željko Agić and Nikola Ljubešić. 2014. The SETimes.HR linguistically annotated corpus of Croatian. In *Proceedings of LREC 2014*.

[Agić et al.2013a] Željko Agić, Nikola Ljubešić, and Danijela Merkler. 2013a. Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57, Sofia, Bulgaria, August. Association for Computational Linguistics.

[Agić et al.2013b] Željko Agić, Danijela Merkler, and Daša Berović. 2013b. Parsing Croatian and Serbian by using Croatian dependency treebanks. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2013)*.

[Baroni et al.2008] Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. 2008. Cleaneval: a competition for cleaning web pages. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

[Baroni et al.2009] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, pages 209–226.

[Bohnet2010] Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *The 23rd International Conference on Computational Linguistics (COLING 2010)*.

[Halácsy et al.2007] Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 209–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Jongejan and Dalianis2009] Bart Jongejan and Hercules Dalianis. 2009. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 145–153.

[Kohlschütter et al.2010] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In Brian D. Davison, Torsten Suel, Nick Craswell, and Bing Liu, editors, *WSDM*, pages 441–450. ACM.

[Ljubešić and Erjavec2011] Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic*, Lecture Notes in Computer Science, pages 395–402. Springer.

[Lui and Baldwin2012] Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *ACL (System Demonstrations)*, pages 25–30.

[Lui et al.2014] Marco Lui, Jey Han Lau, and Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*.

[Schäfer and Bildhauer2012] Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Schäfer et al.2013] Roland Schäfer, Adrien Barbaresi, and Felix Bildhauer. 2013. The good, the bad, and the hazy: Design decisions in web corpus construction. In *Proceedings of the 8th Web as Corpus Workshop (WAC8)*.

[Suchomel and Pomikálek2012] Vít Suchomel and Jan Pomikálek. 2012. Efficient web crawling for large text corpora. In Serge Sharoff Adam Kilgarriff, editor, *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43, Lyon.

[Tiedemann and Ljubešić2012] Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India.