HindEnCorp – Hindi-English and Hindi-only Corpus for Machine Translation

Ondřej Bojar*, Vojtěch Diatka†, Pavel Rychlý‡, Pavel Straňák*
Vít Suchomel‡, Aleš Tamchyna*, Daniel Zeman*

*Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
surname@ufal.mff.cuni.cz
†Charles University in Prague, Faculty of Arts, Department of Linguistics
vojta.diatka@gmail.com
‡Natural Language Processing Centre, Faculty of Informatics, Masaryk University
pary@fi.muni.cz, xsuchom2@fi.muni.cz

Abstract

We present HindEnCorp, a parallel corpus of Hindi and English, and HindMonoCorp, a monolingual corpus of Hindi in their release version 0.5. Both corpora were collected from web sources and preprocessed primarily for the training of statistical machine translation systems. HindEnCorp consists of 274k parallel sentences (3.9 million Hindi and 3.8 million English tokens). HindMonoCorp amounts to 787 million tokens in 44 million sentences.

Both the corpora are freely available for non-commercial research and their preliminary release has been used by numerous participants of the WMT 2014 shared translation task.

Keywords: corpora; parallel corpora; machine translation

1. Introduction

Hindi language is mother tongue of nearly 260 million speakers in India. There are also approximately 160 million speakers fluent in Hindi as their second language. According to the Ethnologue 17[1] it is the fourth largest language in terms of native speakers. It is not spoken only on the Indian subcontinent but also in Nepal, Fiji, Bangladesh etc.

It is quite striking that given this size of Hindi there are very few language resources. Apart from rather small parallel and monolingual corpora (see Section 2.) we weren't able to find any robust ones. We think that such a language deserves a big parallel and monolingual corpus. Since good knowledge of English is not common (according to India Human Development Survey[2] 72% of adult males and 82% adult females in India do not speak English), there is a good opportunity for machine translation from English to Hindi and vice versa.

To meet this need, we collected several already existing resources and added some new. We describe our efforts collecting and cleaning up a Hindi-English parallel corpus and a Hindi monolingual corpus. The resources are primarily aimed at training of statistical machine translation systems, but other uses, such as linguistic analyses, are also possible. We believe that the Hindi monolingual corpus will be one of the largest corpora currently available and it can serve as a very good source for any corpus study of contemporary Hindi as used on the web.

A preliminary version of both of the corpora was used by participants of the WMT[3] shared translation task between English and Hindi.

2. Data Sources

Our current parallel corpus consists of a few parts. Some of them were previously exploited and examined by Bojar et al. (2010), some are new additions to our collection.

The main sources of the current release are summarized in the following sections.

2.1. Parallel Corpora

HindEnCorp parallel texts come from the following sources:

Tides, which contains 50K sentence pairs taken mainly from news articles. This dataset was originally collected for the DARPA-TIDES surprise-language contest in 2002, later refined at IIIT Hyderabad and provided for the NLP Tools Contest at ICON 2008 (Venkatapathy, 2008).

Commentaries by Daniel Pipes contain 322 articles in English written by the journalist Daniel Pipes and translated into Hindi.

EMILLE. This corpus (Baker et al., 2002) consists of three components: monolingual, parallel and annotated corpora. There are fourteen monolingual subcorpora, including both written and (for some languages) spoken data for fourteen South Asian languages. The EMILLE monolingual corpora contain in total 92,799,000 words (including 2,627,000 words of transcribed spoken data

[1] https://www.ethnologue.com/
[2] http://ihds.umd.edu/
[3] http://www.statmt.org/wmt14

for Bengali, Gujarati, Hindi, Punjabi and Urdu). The parallel corpus consists of 200,000 words of text in English and its accompanying translations into Hindi and other languages.

Smaller datasets as collected by Bojar et al. (2010) include the corpus used at ACL 2005 (a subcorpus of EMILLE), a corpus of named entities from Wikipedia (crawled in 2009), and Agriculture domain parallel corpus.

For the current release, we are extending the parallel corpus using these sources:

The Indic multi-parallel corpus (Birch et al., 2011; Post et al., 2012) is a corpus of texts from Wikipedia translated from the respective Indian language into English by non-expert translators hired over Mechanical Turk. The quality is thus somewhat mixed in many respects starting from typesetting and punctuation over capitalization, spelling, word choice to sentence structure. A little bit of control could be in principle obtained from the fact that every input sentence was translated 4 times. We used the 2012 release of the corpus.[4]

Launchpad.net is a software collaboration platform that hosts many open-source projects and facilitates also collaborative localization of the tools. We downloaded all revisions of all the hosted projects and extracted the localization (`.po`) files. Technically, this is a relatively high quality resource with manual sentence alignments (implied by the way the translations were created), but the domain is rather distant from natural text.

TED talks [5] held in various languages, primarily English, are equipped with transcripts and these are translated into 102 languages. There are 179 talks for which Hindi translation is available.

Intercorp (Čermák and Rosen, 2012) as a whole is a large multilingual parallel corpus of 32 languages including Hindi. The central language used for alignment is Czech. Intercorp's core texts amount to 202 million words. These core texts are most suitable for us because their sentence alignment is manually checked and therefore very reliable. They cover predominately short stories and novels. There are seven Hindi texts in Intercorp. Unfortunately, only for three of them the English translation is available; the other four are aligned only with Czech texts.

Other smaller datasets. This time, we added Wikipedia entities as crawled in 2013 (including any morphological variants of the named entity that appears on the Hindi version of the Wikipedia page) and words, word examples and quotes from the Shabdkosh online dictionary.

2.2. Monolingual Hindi Corpora

The second main result of our work is HindMonoCorp, a Hindi monolingual corpus. It is based primarily on web crawls performed using various tools and at various times. Since the web is a living data source, we treat these crawls as completely separate sources, despite the fact that they may overlap. To estimate the magnitude of this overlap, we compared the total number of segments if we concatenate the individual sources (each source being de-duplicated on its own) with the number of segments if we de-duplicate all sources together. The difference is just around 1%, confirming, that various web crawls (or their subsequent processings) differ significantly.
HindMonoCorp contains data from:

SpiderLing is a web crawl carried out during November and December 2013 using SpiderLing (Suchomel and Pomikálek, 2012). The pipeline includes extraction of plain texts and deduplication at the level of documents, see below.

CommonCrawl[6] is a non-profit organization that regularly crawls the web and provides anyone with the data. We are grateful to Christian Buck for extracting plain text Hindi segments from the 2012 and 2013-fall crawls for us.

Hindi web texts (HWT), a monolingual corpus containing mainly Hindi news articles has already been collected and released by Bojar et al. (2008)[7]. We use the HTML files as crawled for this corpus in 2010 and we add a crawl performed in 2013 and re-process them with the current pipeline. These sources are denoted HWT 2010 and HWT 2013 in the following.

Hindi corpora in W2C have been collected by Martin Majliš during his project to automatically collect corpora in many languages (Majliš and Žabokrtský, 2012). There are in fact two corpora of Hindi available[8]—one from web harvest (W2C Web) and one from the Wikipedia (W2C Wiki).

RSS Feeds from Webdunia.com and the Hindi version of BBC International followed by our custom crawler from September 2013 till January 2014.

Intercorp data included in our monolingual corpus contains seven texts in Hindi. Most of them originated in English or Czech but some are original Hindi short stories and novels.

[4] `http://joshua-decoder.org/data/indian-parallel-corpora/`

[5] `http://www.ted.com/`

[7] `http://hdl.handle.net/11858/00-097C-0000-0001-CC1E-B`

[8] `http://ufal.mff.cuni.cz/~majlis/w2c/download.html`

We would like HindMonoCorp to contain also data from open source corpus OPUS[9] or Forum for Information Retrieval Evaluation (FIRE)[10]. Unfortunately, we haven't been able to include these sources into our corpus yet but we plan to do so in the next release.

3. Processing Pipeline

Our processing pipeline begins with the acquisition of the source documents and the extraction of plain text from them. These early stages of processing include encoding and language detection, format stripping and sometimes deduplication. The exact steps and also their order is usually source-dependent, because that way, the most efficient and at the same time convenient method can be applied.

- **Language identification** and **Encoding detection** is very important and not quite trivial in case of Hindi texts, since custom 8bit encodings once prevalent in India (together with custom fonts) are still in use by some news webs. However Unicode data are now widely available and the vast majority of our texts were obtained in this encoding. Detecting the encoding given the language can be efficiently performed by the `chared` tool (Pomikálek, 2011a), which was used for the SpiderLing section of our data. For CommonCrawl, CLD2[11] was used to identify Hindi segments.

  Some data sources naturally contain metadata that allow to reliably identify English and Hindi segments, e.g. TED or Launchpad.

- **Extraction of plain text**, which in our case mainly means writing specific extractors for data sources with a clear and consistent format or the application of a generic HTML stripping tool for the diverse web sources.

  Reliable extraction of plain text with the handcrafted approach pays off only for sources of parallel data, e.g. TED talks.

  For monolingual data, three different HTML strippers were used: W2C used its accompanying tool by Majliš and Žabokrtský (2012). SpiderLing, HWT and RSS Feeds used the jusText tool (Pomikálek, 2011d; Pomikálek, 2011b). The processing of CommonCrawl HTML sources was not ideal: tags were simply removed and the corpus was de-duplicated at the level of lines. In contrast to what web browsers do, the source text of the web page was not reflowed so any line breaks in the middle of the sentence in the source lead to sentences cut too short.

- **De-duplication** is also one of the necessary tasks. We apply it at different processing stages depending on the specifics of the source. The best option is de-duplication at the level of documents, because it best preserves the distribution characteristics of words and phrases.[12] For sources where the documents are difficult to identify or often overlap too much (such as web pages), other methods than the exact identity of documents is desirable.

  The SpiderLing section was processed with onion (Pomikálek, 2011c) that works at the level of documents but considers n-grams of tokens. Of a collection of too similar documents, only one representative is left.

  All other sources except for Intercorp and our parallel data are de-duplicated at the level of segments, thus unfortunately distorting the distributions.

The rest of the pipeline is common to all sources and uses the techniques well tested in the CzEng corpus (Bojar et al., 2012). It consists of the following steps:

- **Sentence segmentation** is carried out by TrTok (Maršík and Bojar, 2012), a trainable tokenizer that was used for processing the English side of CzEng and already contains some training examples for Hindi. We extended the Hindi model for TrTok with a list of manually extracted abbreviations and examples of their use in context and we also double checked and slightly corrected the existing training examples for Hindi sentence segmentation.

- **Sentence alignment** was performed using Hunalign (Varga et al., 2005) in the previous version of our corpus. Other options that we considered included Bleualign (Sennrich and Volk, 2011), which relies on a baseline MT system to translate one of the corpora into the other language, and Gargantua (Braune and Fraser, 2010) which was shown to outperform Hunalign for Urdu-English (Abdul-Rauf et al., 2012) and is known to work better especially for corpora where the sentences do not align 1-to-1 that often. Since the alignment quality of Hunalign seemed acceptable at the first sight, we did not do any experimental comparison with the other tools.

  Note that some sources of HindEnCorp are naturally sentence-segmented and parallel such as the ACL 2005 corpus or the Wikipedia named entities. We skipped automatic sentence segmentation and alignment for these sources.

- **Cleaning and normalization** aims at removing the most apparent typesetting errors in the text. More details are provided in Section 4.

- **Automatic quality checks** significantly increased the quality of the CzEng corpus. We considered them also for HindEnCorp, but in the end,

[9] `http://opus.lingfil.uu.se/`
[10] `http://www.isical.ac.in/~fire/`
[11] `http://code.google.com/p/cld2/`

[12] However, Bojar et al. (2010) have found that Tides and Emille overlap significantly.

there was no benefit from them for the parallel data, see Section 5.

- **Morphological analysis** significantly increases value of the data, both for machine translation and for human use. Previously, we released only tokenised texts, even though in our experiments we had used a Hindi morphological analyzer by Shrivastava and Bhattacharyya (2008). This time around, we are adding also morphologically processed (lemmatized and morphologically tagged) versions of the segmented texts both in the parallel and Hindi-only datasets. For morphological analysis of the English side of the parallel data we employed Morče tagger (Spoustová et al., 2007) available in the Treex platform (Popel and Žabokrtský, 2010)[13], formerly known as TectoMT.

  For Hindi texts, there are several options available. In the past we used the tagger by Shrivastava and Bhattacharyya (2008). Currently there seem to be more viable options including Shallow parser developed by IIIT Hyderabad[14], and Siva Reddy's POS tagger[15]. In the end, we used the latter and provided a wrapper for it into the Treex processing platform.

4. Cleaning and Normalization

Large collections of texts are bound to contain significant amount of noise, e.g. due to varying typesetting conventions and typesetting errors. This is perhaps even more true for languages like Hindi, where the population of Hindi speakers is still in the process of adopting the Internet and the growth in the number of Internet users is still very fast.

We do not have the ambition to tackle noise at the level of words or longer units, as described by Bojar et al. (2010). On the other hand, we at least try to resolve some of the most prominent character-level inconsistencies and errors. Where there is a clear and undisputable automatic correction possible, our script modifies the data.

Aside from removing various non-printable characters and normalizing Unicode to canonical decomposition (Normalization Form D, NFD), we correct the typesetting of Devanagari nukta. This diacritic mark can in Hindi follow only a limited set of characters (क़, ख़, ग़, ज़, ड़, ढ़, फ़, ऴ़) and we remove it elsewhere. We also remove occasional sequences of nuktas which probably serve as a graphical delimiters.

Many phenomena however do not have a solution that one could pick without further disambiguation or deliberate loss of information. For instance, the habits of Hindi writers or web sites differ with respect to indicating the end of the sentence (Devanagari danda "।" vs. plain ASCII full stop "."). A similar variance exists for writing digits in Devanagari ("०१२३४५६७८९") or Western Arabic style ("0123456789"). Similarly, some of our sources were already "excessively" normalized in various ways (period instead of danda, lowercasing of English, tokenization etc.) and reconstructing the proper typesetting is impossible. For these cases, we only report indicative statistics.

Table 1 provides the details about our cleaning and statistics about phenomena that we do not normalize. To ease the comparison of various sources, we report the statistics relative to the sentence count. So if e.g. the danda was seen on average in every sentence from the given source, we report 100.0.

Additionally, we use our script from CzEng to reconstruct paired curly quotes ("") in both Hindi and English using various heuristics.

5. Quality Checks

Based on our experience with CzEng, we expected that the quality of the parallel data can be increased by an ensemble of automatic checks as described in Bojar et al. (2012). The ensemble consists of various language dependent and independent features or indicators that are in the end used in a supervised classifier trained on a few hundred sentence pairs to distinguish parallel and erroneous pairs.

When creating this training data for HindEnCorp, we randomly selected one thousand sentence pairs. We relied on Google Translate[16] to provide translations of the Hindi sentences. We then checked whether the sentences are indeed parallel and whether the English side is correct. We found only 39 erroneous sentence pairs in this data sample, suggesting that over 96% of the corpus are clean, parallel texts. Due to the natural trade-off between precision and recall of the classifier, we concluded that there is no benefit in applying it to HindEnCorp.

6. Corpus Statistics

Tables 2 and 3 report statistics for the final versions of HindEnCorp and HindMonoCorp, respectively.

The number of tokens is reported after tokenization as needed by the two taggers.

7. Availability

HindEnCorp and HindMonoCorp home is here:

`http://ufal.mff.cuni.cz/hindencorp/`

Both the corpora are also easily available for non-commercial use, including research, in the Lindat/Clarin repository:

- `http://hdl.handle.net/11858/00-097C-0000-0023-625F-0` (HindEnCorp 0.5)
- `http://hdl.handle.net/11858/00-097C-0000-0023-6260-A` (HindMonoCorp 0.5)

[13] `http://ufal.mff.cuni.cz/treex`

[14] `http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php`

[15] `http://sivareddy.in/downloads#hindi_tools`

[16] `http://translate.google.com`

	CommonCrawl	SpiderLing	HWT	W2C Web	W2C Wiki	RSS	Launchpad	TIDES	WikiNE	TED	Intercorp	Indic	Emille	DanielPipes	Dictionaries	ACL2005	Agrocorpus
nukta checked	28.1	37.7	37.4	39.6	28.0	35.7	18.1	150.9	14.5	18.6	67.1	26.8	39.6	25.8	21.5	52.6	68.8
after a bad letter dropped	0.2	0.4	0.5	0.3	0.2	0.1	0.1	96.9	0.1	0.0	0.0	0.5	0.0	0.6	0.0	0.0	0.3
after a good letter preserved	27.9	37.3	36.9	39.3	27.8	35.5	18.0	54.0	14.4	18.5	67.1	26.3	39.6	25.2	21.5	52.6	68.5
removed sequences of nuktas	0.0	0.0	0.0	0.0	0.0	0.0	7.8	0.0	0.1	0.1	0.0	0.0	103.1	0.1	9.2	0.0	0.0
danda seen	35.1	66.9	70.8	60.6	57.0	79.6	1.8	0.0	0.0	16.7	84.1	59.4	55.4	100.2	28.8	60.9	65.6
full stop seen	51.5	43.9	31.3	59.7	42.3	19.8	31.2	104.6	0.6	19.6	11.9	22.8	8.3	70.0	20.8	6.9	65.9
double danda seen	0.2	0.1	0.0	0.1	1.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
hindi digit seen	7.4	5.7	1.4	5.7	57.5	0.0	1.5	0.0	9.0	6.1	0.2	59.0	12.4	0.0	6.1	9.9	247.2
eurarabic digit seen	70.4	49.2	53.2	33.9	77.1	73.5	11.2	75.8	9.6	6.5	0.4	49.9	70.6	149.8	11.4	35.4	50.1
nbsp changed to space	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.2	0.0	0.0	0.1	0.0	0.0
removed zero width joiner	0.0	0.0	0.0	0.0	0.0	0.0	3.4	0.0	0.1	0.7	0.0	2.2	0.0	0.0	0.0	0.0	0.0
# lines	20.9M	19.4M	14.4M	2.2M	812.8k	259.5k	66.7k	50.0k	46.0k	39.9k	38.4k	37.7k	10.7k	10.6k	6.5k	3.4k	0.6k

Table 1: Statistics on cleaning and typographical conventions in each source.

Source	Sentences (Parallel)	Tokens English	Tokens Hindi
Tides	50.0k	1.23M	1.31M
Indic	37.7k	0.65M	0.69M
DanielPipes	6.6k	0.53M	0.41M
Launchpad	66.7k	0.47M	0.54M
TED	39.8k	0.35M	0.37M
Emille	8.9k	0.17M	0.16M
Intercorp	7.5k	0.13M	0.15M
Other smaller sources	56.5k	0.25M	0.26M
Total	273.9k	3.76M	3.88M

Table 2: HindEnCorp 0.5 sections and statistics.

	Hindi Sentences	Hindi Tokens
SpiderLing	19.40M	383.50M
CommonCrawl	18.35M	272.03M
HWT2010	2.19M	42.26M
HWT2013	2.11M	38.81M
W2C Web	1.48M	29.84M
W2C Wiki	0.71M	15.58M
RSS	0.21M	4.10M
Intercorp	30.83k	477.22k
Total	44.49M	786.60M

Table 3: HindMonoCorp 0.5 sections and statistics.

These are persistent addresses that should ensure that the data remain at disposal of the research community.

8. Conclusion and Future Work

We presented HindEnCorp and HindMonoCorp in their release version 0.5. The preliminary release 0.1 of these sizeable resources has been already used in the WMT shared translation task.

Our future plans with HindEnCorp and HindMonoCorp include adding further sources, improving the quality of the corpus by various additional filters and checks, and also adding richer automatic linguistic annotation as tools for Hindi become available.

9. Acknowledgment

We are grateful to Christian Buck for providing us with his Hindi extract from CommonCrawl corpus.

The work on this project was supported by the grant FP7-ICT-2011-7-288487 (MosesCore).

This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

10. References

Abdul-Rauf, Sadaf, Fishel, Mark, Lambert, Patrik, Noubours, Sandra, and Sennrich, Rico. (2012). Extrinsic evaluation of sentence alignment systems. In *Workshop on Creating Cross-language Resources for Disconnected Languages and Styles*, pages 6–10, May.

Baker, Paul, Hardie, Andrew, McEnery, Tony, Cunningham, Hamish, and Gaizauskas, Rob. (2002). Emille, a 67-million word corpus of indic languages: Data collection, markup and harmonisation. In *Proc. of LREC 2002*, pages 819–827, Las Palmas, Canary Islands, May. Language Technologies Research Centre.

Birch, Lexi, Callison-Burch, Chris, Osborne, Miles, and Post, Matt. (2011). The Indic multi-parallel corpus. `http://homepages.inf.ed.ac.uk/miles/babel.html`.

Bojar, Ondřej, Straňák, Pavel, and Zeman, Daniel. (2008). English-Hindi Translation in 21 Days. In *Proceedings of the 6th International Conference On Natural Language Processing (ICON-2008) NLP Tools Contest*, Pune, India, December. NLP Association of India.

Bojar, Ondřej, Straňák, Pavel, and Zeman, Daniel. (2010). Data Issues in English-to-Hindi Machine Translation. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*, pages 1771–1777, Valletta, Malta, May. ELRA, European Language Resources Association.

Bojar, Ondřej, Žabokrtský, Zdeněk, Dušek, Ondřej, Galuščáková, Petra, Majliš, Martin, Mareček, David, Maršík, Jiří, Novák, Michal, Popel, Martin, and Tamchyna, Aleš. (2012). The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 3921–3928, Istanbul, Turkey, May. ELRA, European Language Resources Association.

Braune, Fabienne and Fraser, Alexander. (2010). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 81–89, Stroudsburg, PA, USA. Association for Computational Linguistics.

Čermák, František and Rosen, Alexandr. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3):411–427.

Majliš, Martin and Žabokrtský, Zdeněk. (2012). Language richness of the web. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2927–2934, İstanbul, Turkey. European Language Resources Association.

Maršík, Jiří and Bojar, Ondřej. (2012). TrTok: A Fast and Trainable Tokenizer for Natural Languages. *Prague Bulletin of Mathematical Linguistics*, 98:75–85, September.

Pomikálek, Jan. (2011a). Chared, `http://hdl.handle.net/11858/00-097C-0000-000D-F67A-9`.

Pomikálek, Jan. (2011b). jusText, `http://hdl.handle.net/11858/00-097C-0000-000D-F696-9`.

Pomikálek, Jan. (2011c). onion, `http://hdl.handle.net/11858/00-097C-0000-000D-F67B-7`.

Pomikálek, Jan. (2011d). *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis, Ph. D. thesis, Masaryk University.

Popel, Martin and Žabokrtský, Zdeněk. (2010). TectoMT: Modular NLP framework. In Loftsson, Hrafn, Rögnvaldsson, Eirikur, and Helgadottir, Sigrun, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *LNCS*, pages 293–304, Berlin / Heidelberg. Iceland Centre for Language Technology (ICLT), Springer.

Post, Matt, Callison-Burch, Chris, and Osborne, Miles. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada, June. Association for Computational Linguistics.

Sennrich, Rico and Volk, Martin. (2011). Iterative, MT-based sentence alignment of parallel texts . In *NODALIDA 2011, Nordic Conference of Computational Linguistics*, May.

Shrivastava, Manish and Bhattacharyya, Pushpak. (2008). Hindi pos tagger using naive stemming : Harnessing morphological information without extensive linguistic knowledge. In *ICON - 2008 - 6th International Conference on Natural Language Processing*.

Spoustová, Drahomíra, Hajič, Jan, Votrubec, Jan, Krbec, Pavel, and Květoň, Pavel. (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, ACL 2007*, pages 67–74, Praha.

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

Varga, Dániel, Németh, László, Halácsy, Péter, Kornai, András, Trón, Viktor, and Nagy, Viktor. (2005). Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing RANLP 2005*, pages 590–596, Borovets, Bulgaria.

Venkatapathy, Sriram. (2008). NLP Tools Contest – 2008: Summary. In *Proceedings of ICON 2008 NLP Tools Contest*, Pune, India, December. Language Technologies Research Centre.