Rapid creation of large-scale corpora and frequency dictionaries

Attila Zséder[1], Gábor Recski[1], Dániel Varga[2], András Kornai[1]

[1]Computer and Automation Research Institute
Hungarian Academy of Sciences
1111 Budapest, Kende u. 13-17
{zseder,recski,kornai}@sztaki.hu
[2]Media Education and Research Center
Budapest University of Technology and Economics
1111 Budapest, Egry József u. 1.
daniel@mokk.bme.hu

Abstract

We describe, and make public, large-scale language resources and the toolchain used in their creation, for fifteen medium density European languages: Catalan, Czech, Croatian, Danish, Dutch, Finnish, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovak, Spanish, and Swedish. To make the process uniform across languages, we selected tools that are either language-independent or easily customizable for each language, and reimplemented all stages that were taking too long. To achieve processing times that are insignificant compared to the time data collection (crawling) takes, we reimplemented the standard sentence- and word-level tokenizers and created new boilerplate and near-duplicate detection algorithms. Preliminary experiments with non-European languages indicate that our methods are now applicable not just to our sample, but the entire population of digitally viable languages, with the main limiting factor being the availability of high quality stemmers.

Keywords: Web corpus, frequency dictionary, hun* tools

Using the web as a source of linguistic data is by no means a new idea: the first efforts in this direction were made over a decade ago (Resnik 1999, Varantola 2000, Davies 2001), and in 2003 *Computational Linguistics* devoted an entire special issue to the subject (Kilgarriff and Grefenstette 2003). It is all the more surprising that a decade later language resources based on web corpora are, aside from a handful of major languages, still largely missing. Perhaps unsurprisingly, the largest European languages (English, German, and French) fared best, as gigaword corpora and frequency dictionaries have already been created for these as part of the `Wacky` project (Baroni and Kilgarriff 2006, Baroni et al. 2009). The resulting `ukWac`, `deWac`, and `frWac` corpora contain billions of tokens and the derived frequency dictionaries (though not the corpora themselves) are freely available for download. Another discernible tendency of the decade is the rise of national corpora (Tadić 2002, Przepiórkowski et al. 2008, Kucera 2002). Remarkably, the fruits of such state- and EU-funded efforts tend to remain behind query interface walls: individual results pertaining to individual words may be accessible, but the data as a whole is not available for download (see Table 3). Since the fundamental idea is quite sound for medium- and small-density languages as well, we would have expected there to be a significant selection of linguistic data accessed by means of crawling. Yet for most of the medium-size European languages frequency dictionaries derived from gigaword corpora are still unavailable.

Our goal in this paper is to show that a fairly simple pipeline of (generic) crawling and only minimally language-specific

Language	Largest corpus	tokens (M)	Reference	URL
Catalan	CUCWeb	166	Boleda et al. 2006	ramsesii.upf.es/cucweb/
Croatian	Croatian Nat. Corpus	100	Tadic 2002	www.hnk.ffzg.hr/
Czech	Czech National Corpus	1300	Kucera 2002	ucnk.ff.cuni.cz/
Danish	KorpusDK	56	n/a	ordnet.dk/korpusdk_en/
Dutch	Dutch Parallel Corpus	10	Paulussen et al. 2006	www.kuleuven-kortrijk.be/DPC
Finnish	Finnish Text Collection	180	various	www.csc.fi/english/research/software/ftc
Indonesian	SEALang Library	5	n/a	sealang.net/indonesia/corpus.htm
Lithuanian	Corpus of Lithuanian	180	Marcinkevičienè 2004	donelaitis.vdu.lt/
Norwegian	noWaC	700	Guevara 2010	www.tekstlab.uio.no/nowac/
Polish	Polish National Corpus	1200	Przepiórkowski 2008	nkjp.pl
Portuguese	Corpus do Português	45	Davies & Ferreira 2006	www.corpusdoportugues.org
Romanian	Romanian Corpus	50	n/a	www.cse.unt.edu/rada/downloads.html
Serbian	CSL	11	Kostić 2001	www.serbian-corpus.edu.rs/
Slovak	Slovak National Corpus	719	Horák et al. 2004	korpus.juls.savba.sk/
Spanish	Corpus del Espanol	100	Davies 2001	www.corpusdelespanol.org/
Swedish	Korp	910	various	spraakbanken.gu.se/

Table 1: Existing corpora

data cleaning steps can produce useful basic language resources, in particular corpora and frequency dictionaries that are considerably larger than those currently available. Section 1. of this paper describes the crawling process used to obtain raw data and the open-source toolchain we used to process the crawl output. We use this method to create webcorpora for 15 languages, preferring those for which open-source lemmatizers were available, as these allow us to derive frequency dictionaries from the data. The new corpora are presented and compared to existing resources in section 2. All corpora and dictionaries are made available for download at http://hlt.sztaki.hu/resources. Some implications of this work for digitally endangered languages are discussed in the concluding Section 3.

1. Collecting and processing the data

We acquire raw data for each language by crawling the relevant top-level internet domains (cf. Table 4). Since the well-known `Heritrix` crawler (see http://crawler.archive.org) tends to slow down and even halt after several days of operation, which makes significant human supervision essential to its operation, we use the less well known but considerably faster and autonomous `wire` crawler (Castillo et al 2005). In all fairness, what makes `heritrix` slow is precisely its excellence in the task it was designed for, building complete snapshots, and what makes `wire` fast is the radical pruning of sites (e.g. that no timed-out site is tried a second time), a strategy only made possible by the fact that all we need are large samples, completeness is not a goal. In our experience, `wire` will easily sustain 10-20 GB/day throughputs, while `heritrix` will slow down by an order of magnitude after the first day.

Stage	%	Av (GB)	Stdev (GB)
Crawl		97.4	46.4
HTML, boilerplate	100.0	14.2	5.1
Sentence filtering	67.9	9.7	4.0
Language detection	44.8	6.4	3.2
Duplicate filtering	43.5	6.2	3.0
Near-duplicate filt	37.4	5.3	2.4

Table 2: Average data sizes at major stages of the pipeline

The pipeline used in creating the **Wac* corpora is described in detail in the papers of Baroni and his co-workers cited above, but only a few components of the toolchain are publicly accessible. Therefore, we created our own tools (Halácsy et al. 2008), and made sure their runtime is negligible compared to the time it takes to crawl. By now, our tools are capable of processing a week's worth of crawl data in a matter of hours, while the filtering process used in the `Wacky` project is reported to have taken several days for each language. All components of our toolchain are open source (LGPL) and the packaged pipeline is freely downloadable at https://github.com/zseder/webcorpus.

The pipeline takes as its input raw HTML documents. The first steps involve **stripping all HTML data** from documents except the paragraph delimiter `<p>` and **discarding 'boilerplate' sections** (recurring and linguistically irrelevant sections of webpages, such as menus or copyright notices) by identifying paragraphs that occur with greater frequency than a given limit k (currently set at 20) and removing all but the first k instances. Since these steps compress the data seen by later stages of the pipeline drastically (to about 1/7th of the original crawl size), in Table 2 we take the output of these as 100%.

Next we perform **whitespace normalization** and **resolution of HTML character references**, e.g. converting `&#xa3;` to £) before proceeding to **sentence tokenization**.

In order to split our data into sentences efficiently we reimplemented in `flex` the standard sentence-level tokenizer due to Philipp Koehn. In an attempt to improve the quality of our data we discard all sentences which do not end with one of the punctuation marks `.,:?!` and all documents which do not contain at least three sentences. These steps again reduce data size by about 30%.

The next major step of **cleaning** our corpora was to pass them through the `hunspell` spellchecker (Németh et al. 2004) and discard documents for which the ratio of unrecognized words was above some threshold (set at 60% based on our earlier work on Hungarian). This step not only improves the overall quality of documents in a corpus of a given language but is also a means of language-detection: documents written in a language other than that of the corpus are also discarded at this point. Since no `hunspell` dictionary is available for the Finnish language, we used Gertjan van Noord's n-gram language classifier `textcat` (see http://www.let.rug.nl/∼vannoord/TextCat) to detect Finnish data in our crawl of the `.fi` domain.

Next we detect **duplicate webpages** using hash-based comparison and keep only the first occurrence of each document. In order to detect **near-duplicate** documents as well, we use the `shash` C implementation of the similarity hash method (see https://github.com/vilda/shash, Charikar 2002, Manku et al. 2007). On the average, only about 5-6% of the original crawl remains.

2. Corpora and frequency dictionaries

Using the pipeline described above we created corpora for 15 European languages: Catalan, Croatian, Czech, Danish, Dutch, Finnish, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovak, Spanish and Swedish. As an experiment with crawling a remote section of the web, we also created resources for Bahasa Indonesia. Since the results seem reasonable, we we are making these available as well – we return to this matter in the concluding Section 3.

For these languages, Table 1 shows the size of the largest existing corpora that we know of. Table 3 shows the availability of these resources. For several languages the largest searchable corpora contain hundreds of millions of tokens. However, none of these are freely available for download. Our toolchain produced corpora with sizes in the hundred millions, see Table 4.

The next step involved counting the number of occurrences of each token and thus creating a frequency dictionary of all word forms of a language that are present in the data gathered. Since it is the frequency of a lemma, rather than an individual word form, that is relevant for most purposes, we passed each dictionary through `hunspell`

corpus	download	search
CUCWeb	no	yes
Czech National Corpus	no	yes
Croatian National Corpus	no	yes
KorpusDK	no	yes
Dutch Parallel Corpus	no	no
Finnish Text Collection	some	yes
SEALang Library	no	yes
Corpus of Lithuanian	no	yes
noWaC	no	yes
Polish National Corpus	no	yes
Corpus do Português	no	yes
Romanian Corpus	no	no
CSL	no	no
Slovak National Corpus	no	yes
Corpus del Español	no	yes
Korp	some	yes

Table 3: Access to preexisting corpora

Language	domain	tok	lem	unk	ratio
Catalan	cat	658	64	215	32.8
Czech	cz	295	179	75	25.4
Croatian	hr	1491	202	538	36.1
Danish	dk	492	144	148	30.1
Dutch	nl	1989	104	634	31.9
Finnish	fi	395	107	153	38.7
Indonesian	id	310	20	119	38.3
Lithuanian	lt	1405	59	600	42.7
No (Bokmål)	no	1620	720	520	32.1
No (Nynorsk)	no	26	106	6.7	25.6
Polish	pl	274	986	104	37.8
Portuguese	pt	963	32	302	31.4
Romanian	ro	1067	75	437	40.8
Serbo-Croatian	rs+hr	2337	340	298	35.3
Serbian	rs	845	201	836	35.8
Slovak	sk	862	148	315	36.5
Spanish	es	1397	53	433	31.0
Swedish	se	893	344	280	31.4

Table 4: Main parameters of the newly created corpora: (unknown) tokens in millions, lemmas in thousands, ratio of unknown tokens to total in %.

for lemmatization. For each language `hunspell` failed to recognize some fraction of all word forms, partly due to noise in our data and partly to the incompleteness of individual `hunspell` dictionaries. It is difficult to estimate the relative weight of these two factors, but the proportion of such word forms, listed in Table 4, is more indicative of the coverage of the stemmer than of the quality of the filtered corpus. In the case of Finnish (for which no `hunspell` dictionary is available) we used the open-source FST-based morphological analyzer `omorfi` (see http://gna.org/projects/omorfi).

In average, `hunspell` found some 11% percent of all word forms ambiguous and returned multiple stems. The frequency dictionaries we created contain two figures for each stem: one was obtained by choosing the shortest possible stem for each ambiguous form, while the other is the result of summing the frequencies of all tokens for which the given stem was among the options returned by `hunspell`. Both the frequency dictionaries and the tokenized corpora are available for download at http://hlt.sztaki.hu/resources.

3. Further directions

Thousands of languages are 'modern' in the technical sense of currently being used for day to day communication, but less than half of these have significant literacy, and even the existence of a broad indigenous literary tradition is no guarantee of a standardized orthography. Yet it is virtually impossible to create a significant digital community without common spelling, and entirely impossible to create a communal knowledge repository such as a wikipedia without a vibrant digital community. By this simple criterion, well over 95% of modern languages are digitally endangered, and it is highly unlikely that more than two hundred languages will ever make the transition to the web. In fact, even languages with significant and growing wikipedias, such as Basque (over 120k articles), Irish (14k articles), or Karakalpak (.5k articles), may end up being monuments of digitally moribund languages, as long as they are maintained by small bands of enthusiasts but have practically no other machine readable material.

The standard criterion for the viability of a language (Krauss 2007) is to consider whether children will be speaking it a hundred years hence. If we convert this criterion to the digital age, and admit quite frankly that we have no easy way of predicting usage trends in social media for ten, let alone a hundred years, we are left with a more operational, but rather stark, definition: a language can transition to the digital age only to the extent it produces *new, publicly available* digital material. In this regard, the national corpora-building efforts that neither preserve nor produce publicly available resources (see Table 3) are missing the mark. A lively virtual community of just a few thousand people, each writing just a few hundred words per day, will easily create a quarter billion words in a year, and a gigaword corpus in 3-4 years. Being in use in a wide variety of settings is already an established (pre-digital) criterion of viability (Dorian 1980), and there can be little doubt that a digitally viable language must also boast of several virtual communities engaged in different activities.

The larger aim of our work is to pave the way for creating gigaword corpora not just for a sample, but the entire *population* of digitally viable languages. For most of the languages discussed here, this task is now trivial, and the only limiting factor is the public availability of high quality lemmatizers. The case of Serbo-Croatian (SRC) is rather telling: we could *collect* a good amount of Serbian (Cyrillic SRC) but we can *process* only Croatian (Latin SRC) for want of a Serbian lemmatizer. Another telling example is Nynorsk, which has produced a corpus of only a few megabytes, only 2-3% of the Bokmål corpus, in spite of the fact that the Nynorsk wikipedia (80k articles) is almost a quarter of the Bokmål (330k articles). The conclusion seems inevitable: Nynorsk will join Classical Chinese, Latin, and Sanskrit as a culture-bearing resource that is restricted to enthusiasts much like Klingon or Volapük, but without truly making the transition to the digital age. In

sharp contrast to this, Indonesian, a language that didn't even have a wikipedia ten years ago, has by now a robust enough web presence to enable collection of a significant (310 m words) corpus.

In future work, we plan on evaluating the entire candidate set to see which languages have a digital future, but the lesson from the present work is already clear: in order to guarantee the viability of some digitally endangered language, one needs digital literacy in that language. Publicly available word-level tools, such as spellcheckers, stemmers, and morphological analyzers (all supported by free and open source software in the hunspell framework) are a necessary precondition of digital literacy, and thus, of survival.

Acknowledgements

Zséder implemented the new toolchain, Recski handled the language-specific steps, Varga contributed the original tools, Kornai advised.

4. References

M. Baroni and A. Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 87–90. Association for Computational Linguistics.

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

G. Boleda, S. Bott, R. Meza, C. Castillo, T. Badia, and V. López. 2006. Cucweb: a catalan corpus built from the web. In *Proceedings of the 2nd International Workshop on Web as Corpus*, pages 19–26. Association for Computational Linguistics.

C. Castillo and R. Baeza-Yates. 2005. Wire: an open source web information retrieval environment. *OSWIR 2005*, page 27.

M.S. Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388. ACM.

M. Davies and M. Ferreira. 2006. Corpus do português (45 milhões de palavras, sécs. xiv-xx).

M. Davies. 2001. Creating and using multi-million word corpora from web-based newspapers. *Corpus Linguistics in North America*, pages 58–75.

Nancy C. Dorian. 1980. Language shift in community and individual: The phenomenon of the laggard semi-speaker. *International Journal of the Sociology of Language*, 25:85–94.

E. Guevara. 2010. Nowac: a large web-based corpus for norwegian. In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, pages 1–7. Association for Computational Linguistics.

P. Halácsy, A. Kornai, P. Németh, and D. Varga. 2008. Parallel creation of gigaword corpora for medium density languages–an interim report. In *Proceedings of Language Resources and Evaluation Conference (LREC08). European Language Resources Association*. Citeseer.

A. Horák, L. Gianitsová, M. Šimková, M. Šmotlák, and R. Garabík. 2004. Slovak national corpus. In *Text, Speech and Dialogue*, pages 89–93. Springer.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, pages 333–348.

D. Kostić. 2001. Quantitative description of Serbian language structure. *Institute for Experimental Phonetic and Speech Pathology and Laboratory for Experimental Psychology, Belgrade (in Serbian)*.

M.E. Krauss. 2007. Mass language extinction and documentation: The race against time. In O. Miyaoka, O. Sakiyama, and M.E. Krauss, editors, *The vanishing languages of the Pacific rim*, pages 3–27. Oxford University Press, USA.

K. Kucera. 2002. The Czech National Corpus: Principles, design, and results. *Literary and linguistic computing*, 17(2):245.

G.S. Manku, A. Jain, and A. Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150. ACM.

R. Marcinkeviciene, A. Bielinskiene, V. Daudaravicius, and E. Rimkute. 2004. Corpora for lithuanian language technologies. In *Proceedings of the First Baltic Conference Human Language Technologies*, pages 21–24.

L. Németh, V. Trón, P. Halácsy, A. Kornai, A. Rung, and I. Szakadát. 2004. Leveraging the open-source ispell codebase for minority language analysis. *Proceedings of SALTMIL*, pages 56–59.

H. Paulussen, L. Macken, J. Trushkina, P. Desmet, and W. Vandeweghe. 2006. Dutch parallel corpus: a multi-functional and multilingual corpus. *Cahiers de l'Institut de Linguistique de Louvain*, 32(1-4):269.

A. Przepiórkowski, R.L. Górski, B. Lewandowska-Tomaszczyk, and M. Łazinski. 2008. Towards the National Corpus of Polish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC*.

Philip Resnik. 1999. Mining the web for bilingual text. *Proc 37th ACL*.

M. Tadić. 2002. Building the Croatian National Corpus. In *LREC2002 Proceedings, Las Palmas, ELRA, Pariz-Las Palmas*, volume 2, pages 441–446.