# A Survey on Large Scale Web Based Corpora

Michal Ptaszynski † Rafal Rzepka ‡ Kenji Araki ‡ Yoshio Momouchi §

† High-Tech Research Center, Hokkai-Gakuen University
ptaszynski@hgu.jp

‡ Graduate School of Information Science and Technology, Hokkaido University
{kabura,araki}@media.eng.hokudai.ac.jp

§ Department of Electronics and Information Engineering,
Faculty of Engineering, Hokkai-Gakuen University
momouchi@eli.hokkai-s-u.ac.jp

**Abstract**

Recently there have been several initiatives to create locally accessible large scale corpora based on the contents of the Internet. In this paper we present a survey on several such corpora created for different languages. We compare their distinctive features and the amount of additional annotations provided by the developers of those corpora.

## 1 Introduction

Text corpora are vital linguistic resources in natural language processing. Recent development in technology, one of the symptom of which is frequent switching to 64bit machines and operating systems, has allowed for compilation and efficient processing of billion-word and larger corpora. There have been several initiatives to create large scale corpora, and the need for those is constantly growing. In this paper we describe some of those initiatives. In the following sections we firstly address the question whether it is necessary to create corpora of that size. Next, we describe some of such corpora, compare their scale, features and the amount of additional annotations performed on them. We also dedicate a separate section to compare corpora in the Japanese language. Finally, we conclude the paper and list up some of the possible applications of large corpora.

### 1.1 Size Does Matter

The notion of a "large scale corpus" has appeared in linguistic and computational linguistic literature for many years. However, study of the literature shows that what was considered as "large" ten years ago does not exceed a 5% (border of statistical error) when compared to present corpora. For example, Sasaki et al. [1] in 2001 reported a construction of a question answering (QA) system based on a large scale corpus. The corpus they used consisted of 528,000 newspaper articles. YACIS, one of the corpora described here, consists of 12,938,606 documents (blog pages). The rough estimation indicates that the corpus of Sasaki et al. covers less than 5% of YACIS (in particular 4.08%). In this survey we mostly focused on research scaling the meaning of "large" up to around billion words and more.

It could be questioned whether billion-word and larger corpora are of any use to linguistics and in what sense it is better to use a large corpus than a medium sized one. This question has been answered by most of the researchers involved in the creation of large corpora, thus we will answer it briefly referring to the relevant literature. Baayen [2] notices that language phenomena (such as probability of appearance of certain words within a corpus) are distributed in accordance with Zip's Law. The Zip's Law was originally proposed by George Kingsley Zipf in late 1930's to 1940's [3, 4], who formulated a wide range of linguistic phenomena based on probability. One such phenomenon says that the number of occurrences of words within a corpus decreases in a quadratic-like manner. For example, when all unique words in a corpus are represented in a list with decreasing occurrences, the second word on the list will have a tendency to appear two times less often than the first one. This means that if a corpus is not big enough, many words will not appear in it at all. Baroni and Ueyama [5] and Pomikálek et al. [6] indicate that Zipf's Law is one of the strongest reasons to work with large-scale corpora, if we are to understand the most of the language phenomena and provide statistically reliable proofs for them. There are opponents of uncontrolled over-scaling of corpora, such as Curran (with Osborne in [7]), who show that convergence behavior of words in a large corpus does not necessarily appear for all words and thus it is not the size of the corpus that matters, but the statistical model applied in the processing. However, they do admit that the corpus scale is one of the features that should be addressed in the corpus linguistic research and eventually join the initiative of developing a 10 billion word corpus of English (see Liu and Curran [8]).

Table 1: Comparison of different Web-based corpora, ordered by size (number of words/tokens).

| corpus name | scale (in words) | language | domain | annotation |
|---|---|---|---|---|
| Liu&Curran [8] | 10 billion | English | whole Web | tokenization; |
| YACIS [15] | 5.6 billion | Japanese | Blogs (Ameba) | tokenization, POS, lemma, dependency parsing, NER, affect (emotive expressions, valence, activation, etc.); |
| BiWeC [6] | 5.5 billion | English | whole Web (.uk and .au domains) | POS, lemma; |
| ukWaC | 2 billion | English | whole Web (.uk domain) | POS, lemma; |
| PukWaC (Parsed-ukWaC) [12] | 2 billion | English | whole Web (.uk domain) | POS, lemma, dependency parsing; |
| itWaC [5, 12] | 2 billion | Italian | whole Web (.it domain) | POS, lemma; |
| Gigaword [13] | 2 billion | Hungarian | whole Web (.hu domain) | tokenization, sentence segmentation; |
| deWaC [12] | 1.7 billion | German | whole Web (.de domain) | POS, lemma; |
| frWaC [12] | 1.6 billion | French | whole Web (.fr domain) | POS, lemma; |
| National Corpus of Polish [14] | 1 billion | Polish | multi-domain (newspapers, literature, Web, etc.) | POS, lemma, dependency parsing, named entities, word senses; |

# 2 Large-Scale Corpora

Liu and Curran [8], followed by Baroni and Ueyama [5], indicate at least two types of research dealing with large-scale corpora. One is using popular search engines, such as Google[1] or Yahoo![2]. The second one is crawling the Web and downloading the corpus locally for further annotations and processing.

## 2.1 Search Engine Querying

In research based on querying search engines one gathers estimates of hit counts for certain keywords to perform statistical analysis, or wider contexts of the keywords, called "snippets" (a short, three line long set of text containing the keyword), to perform further analysis of the snippet contents. This refers to what has generally developed as the "Web mining" field. One of the examples is the research by Turney and Littman [9]. They claim to perform sentiment analysis on a hundred-billion-word corpus. By the corpus they mean roughly estimated size of the web pages indexed by AltaVista search engine[3]. However, this kind of research is inevitably constrained with limitations of the search engine's API. Pomikálek et al. [6] indicate a long list of such limitations. Some of them include: limited query language (e.g. no search by regular expressions), query-per-day limitations (e.g. Google allows only one thousand queries per day for one IP address, after which the IP address is blocked - an unacceptable limitation for linguistic research), search queries are ordered with a manner irrelevant to linguistic research, etc.. Kilgariff [10] calls uncritical relying on search engine results a "Googleology" and points out a number of problems search engines will never be able to deal with (such as duplicated documents). Moreover, only Google employees have unlimited and extended access to the search engine results. Kilgariff also proposes an alternative, building large-scale corpora locally by crawling the Web, and argues that it is the optimal way of utilizing the Internet contents for research in linguistics and computational linguistics.

## 2.2 N-gram Based Corpora

There have been several initiatives to build billion-word-scale corpora for different languages. Google is a company that holds presumably the largest text collection in the world. The scale makes it impossible to control, evaluate and fully annotate, which makes it a large collection not fully usable for researchers [10, 6]. However, Google has presented two locally accessible large corpora. One is the "Web 1T (trillion) 5 gram" corpus [11] published in 2006. It is estimated to contain one trillion of tokens gathered from 95 billion sentences. Unfortunately, the contents available for users are only n-grams, from 1 (unigrams) to 5 (pentagrams). The corpus was not processed in any way except tokenization. Also, the original sentences are not available. This makes the corpus, although unmatchable when it comes to statistics of short word sequences, not interesting for language studies, where a word needs to be processed in its context (a sentence, a paragraph, a document). The second one is the "Google Books 155 Billion Word Corpus"[4] announced in 2011. It contains 1.3 million books published between 1810 and 2009 and processed with OCR. This corpus has a larger functionality, such as part of speech annotation and lemmatization of words. However, it is available only as an online interface with a daily access limit per user (1000 queries). The tokenized-only version of the corpus is available, also for several other languages[5], unfortunately only in the n-gram form (no context larger than 5-gram).

[1] http://www.google.com

[2] http://www.yahoo.com

[3] In 2004 AltaVista (http://www.altavista.com/) has become a part of Yahoo!.

[4] http://googlebooks.byu.edu/

[5] http://books.google.com/ngrams/datasets

Table 2: Detailed comparison of different Japanese corpora, ordered by size (number of words/tokens).

| corpus name | scale (in words) | number of documents (Web pages) | number of sentences | size (uncompressed in GB, text only, no annotation) | domain |
|---|---|---|---|---|---|
| YACIS [15] | 5,600,597,095 | 12,938,606 | 354,288,529 | 26.6 | Blogs (Ameba); |
| JpWaC [18] | 409,384,411 | 49,544 | 12,759,201 | 7.3 | Whole Web (11 domains within .jp); |
| jBlogs [5] | 61,885,180 | 28,530 | [not revealed] | .25 (compressed) | Blogs (Ameba,Goo,Livedoor,Yahoo!); |
| KNB [22] | 66,952 | 249 | 4,186 | 450 kB | Blogs (written by students); |

## 2.3 Web-Crawled Corpora

Among corpora created with Web crawling methods, Liu and Curran [8] created a 10-billion-word corpus of English. Although the corpus was not annotated in any way, except tokenization, differently to Google's corpora it is sentence based, not n-gram based. Moreover, it successfully proved its usability in standard NLP tasks such as spelling correction or thesaurus extraction.

The **WaCky** (**W**eb **a**s **C**orpus **k**ool **y**nitiative) [5, 12] project started gathering and linguistically processing large scale corpora from the Web. In the years 2005-2007 the project resulted in more than five collections of around two billion word corpora for different languages, such as English (ukWaC), French (frWaC), German (deWaC) or Italian (itWaC). The tools developed for the project are available online and their general applicability is well established. Some of the corpora developed within the project are compared in table 1.

**BiWeC** [6], or **Bi**g **We**b **C**orpus has been collected from the whole Web contents in 2009 and consists of about 5.5 billion words. The authors of this corpus aimed to go beyond the border of 2 billion words set by the WaCky initiative[6] as a borderline for corpus processing feasibility for modern (32-bit) software.

Billion-word scale corpora have been recently developed also for less popular languages, such as Hungarian [13] or Polish [14].

## 2.4 Japanese Web-Based Corpora

As for corpora in Japanese, despite the fact that Japanese is a well recognized and described world language, there have been only a few corpora of a reasonable size.

**YACIS** or **Y**et **A**nother **C**orpus of **I**nternet **S**entences was collected automatically by Maciejewski et al. [15] from the pages of Ameba blog service. It contains 5.6 billion words within 350 million sentences. It has been annotated with different types of annotations. Ptaszynski et al. [16] annotated it with syntactic information such as POS tagging, lemma, dependency parsing, etc., and Ptaszynski et al. [17] added affective annotations such as emotive expressions, emotion classes, valence, etc..

Srdanović Erjavec et al. [18] used WAC (Web As Corpus) Toolkit[7], developed under the WaCky initiative to gather **JpWaC**, a 400 million word corpus of Japanese. Although JpWac covers only about 7% of YACIS (400 mil. vs 5.6 bil. words), the research is worth mentioning, since it shows that freely available tools developed for European languages are to some extend applicable also for languages of completely different typography, like Japanese[8]. However, they faced several problems. Firstly, they had to normalize the character encoding for all web pages[9] (Ameba blog service, on which YACIS was based, is encoded by default in Unicode). Moreover, since they did not specify the domain, but based the corpus on the whole Web contents, they were unable to deal ideally with the web page metadata, such as the page title, author, or creation date, which differs between domains (Ameba has clear and stable meta-structure).

Baroni and Ueyama [5] developed **jBlogs**, a medium-sized corpus of Japanese blogs containing 62 million words. They selected four popular blog services (Ameba, Goo, Livedoor, Yahoo!) and extracted nearly 30 thousand blog documents. Except part of speech tagging, which was done by a Japanese POS tagger ChaSen, the whole procedure and tools they used were the same as the ones developed in WaCky. In the detailed manual analysis of the jBlogs, Baroni and Ueyama noticed that blog posts contained many Japanese emoticons, namely *kaomoji*. They reported that ChaSen is not capable of processing them, and separates each character adding a general annotation tag ”symbol”. This results in an overall bias in distribution of parts of speech, putting symbols as the second most frequent (nearly 30% of the whole jBlogs corpus) tag, right after ”noun” (about 35%). They considered the frequent appearance of emoticons a major problem in processing blog corpora. In our research we dealt with this problem. To process emoticons we used CAO, a system for detailed analysis of Japanese emoticons developed previously by Ptaszynski et al. [19].

Apart from the above Kawahara and Kurohashi [20] claim the creation of a large, about two-billion-word corpus. However, detailed description of this corpus is not available. Okuno and Sasano from Yahoo! Japan report on developing a large scale blog corpus, similar in form to the Google “Web 1T 5 gram” with only n-grams available for processing [21]. No information on the corpus is yet available except methods of development, tools (tokenization by MeCab, a POS tagger for Japanese) and its

[6] http://wacky.sslmit.unibo.it/

[7] http://www.drni.de/wac-tk/

[8] languages like Chinese, Japanese or Korean are encoded using 2-bite characters.

[9] Japanese can be encoded in at least four standards: JIS, Shift-JIS, EUC, and Unicode.

size (1TB).

Finally, a smaller scale corpus of Japanese blogs, similar to YACIS in the amount and diversification of annotated information, has been developed by Hashimoto et al. in 2010 and published in 2011 [22]. The corpus was developed jointly by the National Institute of Information and Communications Technology, Kyoto University, and the NTT Communication Science Laboratories. The **KNB**[10] corpus contains about 67 thousand words in 249 blog articles. Although it is not a large scale corpus (0.12% of YACIS compared by words/tokens), it developed a certain standard for preparing corpora, especially blog corpora for sentiment and affect-related studies in Japan. The corpus contains all relevant syntactic and morphological annotations, including POS tagging, dependency parsing or Named Entity Recognition. It also contains sentiment-related information. Words and phrases expressing emotional attitude were annotated by laypeople as either positive or negative. One disadvantage of the corpus, except its small scale, is the way it was created. Eighty one students were employed to write blogs about different topics especially for the need of this research. It could be argued that since the students knew their blogs will be read mostly by their teachers, they selected their words more carefully than they would on their private blogs.

## 3 Conclusions

In this paper we performed a survey on Web-based corpora, with a focus on large-scale corpora containing billion words or more. We compared some of their features and the amount of annotations. Additionally we compared some of the Web-based corpora for the Japanese language. There are many applications large corpora could be helpful with. For example, Liu and Curran [8] used their 10-billion-word corpus for tasks such as spelling correction and thesaurus extraction. Ptaszynski et al. [19] used YACIS to evaluate a system for affect analysis of emoticons. Turney and Littman [9] showed that large scale corpora could be useful in research on sentiment analysis. Finally, large corpora can also be applied to creating more detailed sub-corpora for a focused study, or serve as an alternative for systems relying on constant search engine querying (e.g. chatbots).

### Acknowledgments

This research was supported by (JSPS) KAKENHI Grant-in-Aid for JSPS Fellows (Project Number: 22-00358).

## References

[1] Sasaki, Y., Isozaki, H., Taira, H., Hirao, T., Kazawa, H., Suzuki, J., Kokuryo, K., Maeda, E. 2001. "SAIQA : A Japanese QA System Based on a Large-Scale Corpus" [in Japanese], *IPSJ SIG Notes* 2001(86), pp. 77-82.

[2] Baayen, H. 2001. *Word Frequency Distributions*. Dordrecht, Kluwer.

[3] Zipf, George K. 1935. *The Psychobiology of Language*. Houghton-Mifflin.

[4] Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

[5] Baroni, M. and Ueyama, M. 2006. "Building General- and Special-Purpose Corpora by Web Crawling", In *Proceedings of the 13th NIJL International Symposium on Language Corpora: Their Compilation and Application*.

[6] Pomikálek, J., Rychlý, P. and Kilgarriff, A. 2009. "Scaling to Billion-plus Word Corpora, Advances in Computational Linguistics", Advances in Computational Linguistics, *Research in Computing Science*, 41, pp. 3-14.

[7] Curran, J. R. and Osborne, M. 2002. "A very very large corpus doesn't always yield reliable estimates", In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)*, pp. 126-131.

[8] Liu V. and Curran, J. R. 2006. "Web Text Corpus for Natural Language Processing", In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 233-240.

[9] Turney, P. D. and Littman, M. L. 2002. "Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus", National Research Council, Institute for Information Technology, Technical Report ERB-1094. (NRC #44929).

[10] Kilgarriff, A. 2006. "Googleology is Bad Science", Last Words in: *Computational Linguistics*, Vol. 33, No. 1.

[11] Brants, T. and Franz, A. 2006. "Web 1T 5-gram Version 1", Linguistic Data Consortium, Philadelphia.

[12] Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E. 2008. "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora", Kluwer Academic Publishers, Netherlands.

[13] Halacsy, P., Kornai, A., Nemeth, L., Rung, A., Szakadat, I. and Tron, V. 2004. "Creating open language resources for Hungarian". In *Proceedings of the LREC*, Lisbon, Portugal, 2004.

[14] Głowińska, K., Przepiórkowski, A. 2010. "The Design of Syntactic Annotation Levels in the National Corpus of Polish", In: *LREC Proceedings*. http://nkjp.pl/

[15] Maciejewski, J., Ptaszynski, M., Dybala, P. 2010. "Developing a Large-Scale Corpus for Natural Language Processing and Emotion Processing Research in Japanese", In *Proceedings of the International Workshop on Modern Science and Technology (IWMST)*, pp. 192-195.

[16] Ptaszynski, M., Rzepka, R., Araki, K. and Momouchi, Y. 2012. "Annotating Syntactic Information on 5.5 Billion Word Corpus of Japanese Blogs", In *Proceedings of The 18th Annual Meeting of The Association for Natural Language Processing (NLP-2012)*.

[17] Ptaszynski, M., Dybala, P., Rzepka, R., Araki, K. and Momouchi, Y. 2012. "Annotating Affective Information on 5.5 Billion Word Corpus of Japanese Blogs", In *Proceedings of The 18th Annual Meeting of The Association for Natural Language Processing (NLP-2012)*.

[18] Erjavec, I. S., Erjavec, T., Kilgarriff, A. 2008. "A web corpus and word sketches for Japanese", *Information and Media Technologies*, 3(3), pp. 529-551.

[19] Ptaszynski, M., Maciejewski, J., Dybala, P., Rzepka, R., Araki, K. 2010. "CAO: Fully Automatic Emoticon Analysis System", In *Proc. of the 24th AAAI Conference on Artificial Intelligence (AAAI-10)*, pp. 1026-1032.

[20] Kawahara, D. and Kurohashi, S. 2006. "A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis", *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pp. 176-183.

[21] Okuno Y. and Sasano M. 2011. "Language Model Building and Evaluation using A Large-Scale Japanese Blog Corpus" [in Japanese], *The 17th Annual Meeting of The Association for Natural Language Processing*, pp. 955-958.

[22] Hashimoto, C., Kurohashi, S., Kawahara, D., Shinzato, K. and Nagata, M. 2011. "Construction of a Blog Corpus with Syntactic, Anaphoric, and Sentiment Annotations" [in Japanese], *Journal of Natural Language Processing*, Vol. 18, No. 2, pp. 175-201.

[10] Abbreviation of **K**yoto University and **N**TT Lab **B**log Corpus.