北海学園大学
Hokkai-GakuenUniversity

JAPAN SOCIETY FOR THE PROMOTION OF SCIENCE · FOUNDED 1932 ·

A Survey on Large Scale Web Based Corpora

Michal Ptaszynski [1], Rafal Rzepka [2], Kenji Araki [2], Yoshio Momouchi [3]

1) JSPS Research Fellow / Hokkai-Gakuen University, High-Tech Research Center

2) Hokkaido University, Graduate School of Information Science and Technology

3) Hokkai-Gakuen University, Department of Electronics and Information Engineering

A Survey on Large Scale Web Based Corpora

Michal Ptaszynski [1], Rafal Rzepka [2], Kenji Araki [2], Yoshio Momouchi [3]

1) JSPS Research Fellow / Hokkai-Gakuen University, High-Tech Research Center

2) Hokkaido University, Graduate School of Information Science and Technology

3) Hokkai-Gakuen University, Department of Electronics and Information Engineering

Presentation Outline

- Introduction
  - How large are large corpora?
  - Do we need large corpora?
- Research on Large Scale Corpora
  - Search Engine Querying
  - N-gram based corpora
  - Web-crawled corpora
  - Japanese Web-based corpora
- Conclusions

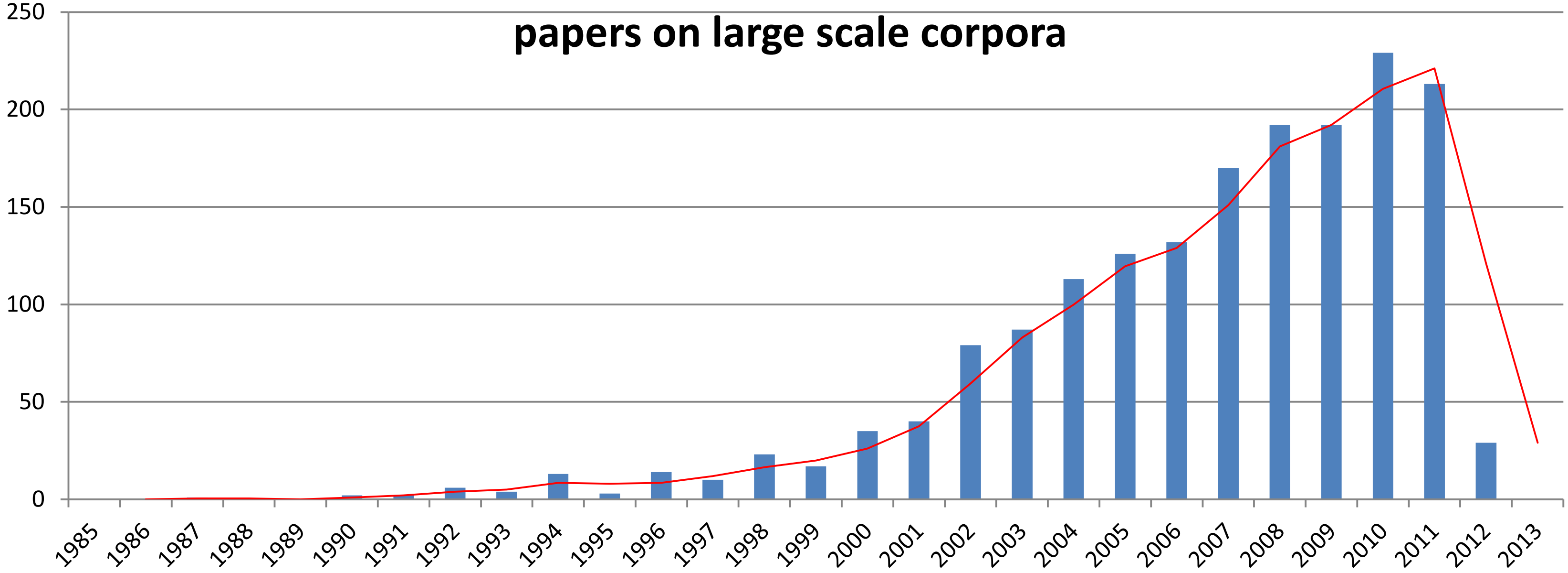
Introduction

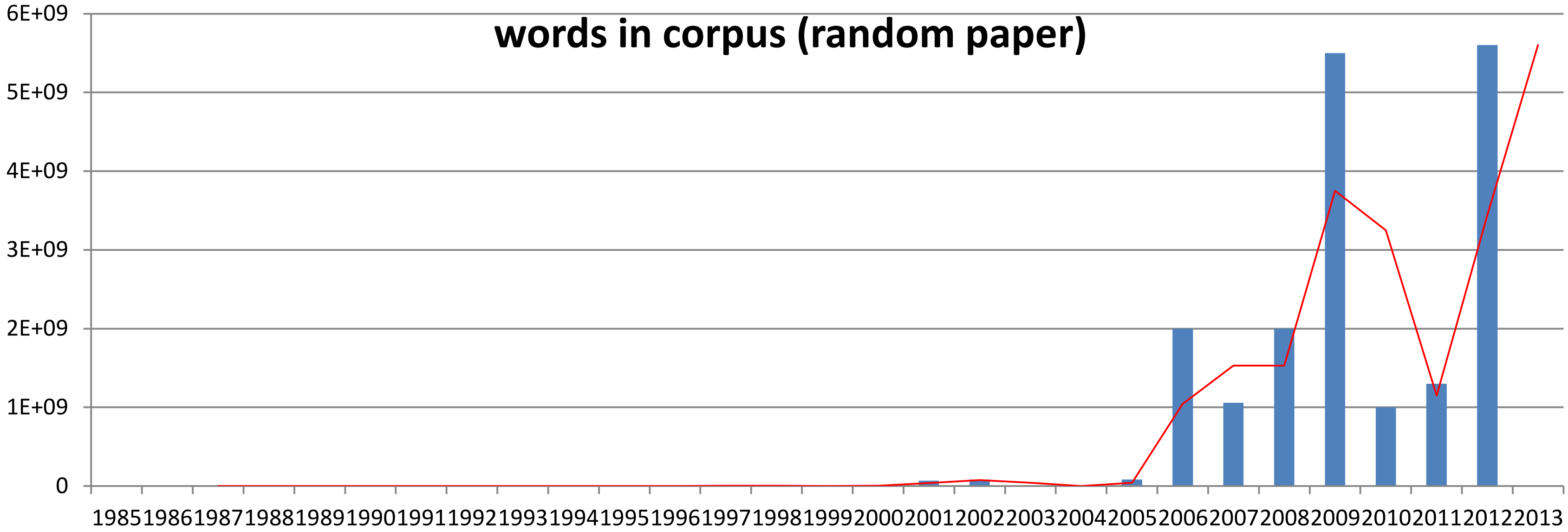
How large are large corpora?

- The notion of a “large scale corpus” has appeared in linguistic and computational linguistic literature for many years.
  - (perhaps) first use of phrase “large scale corpus” 1987

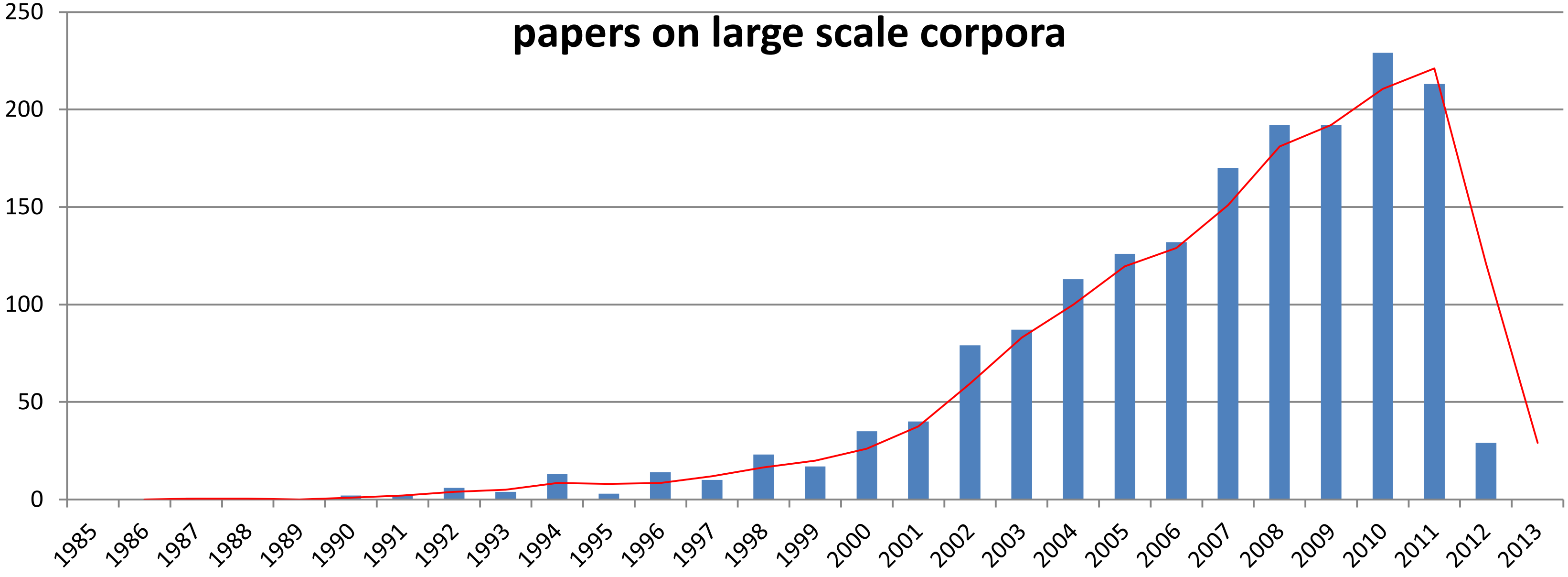
How large are large corpora?

- The notion of a “large scale corpus” has appeared in linguistic and computational linguistic literature for many years.

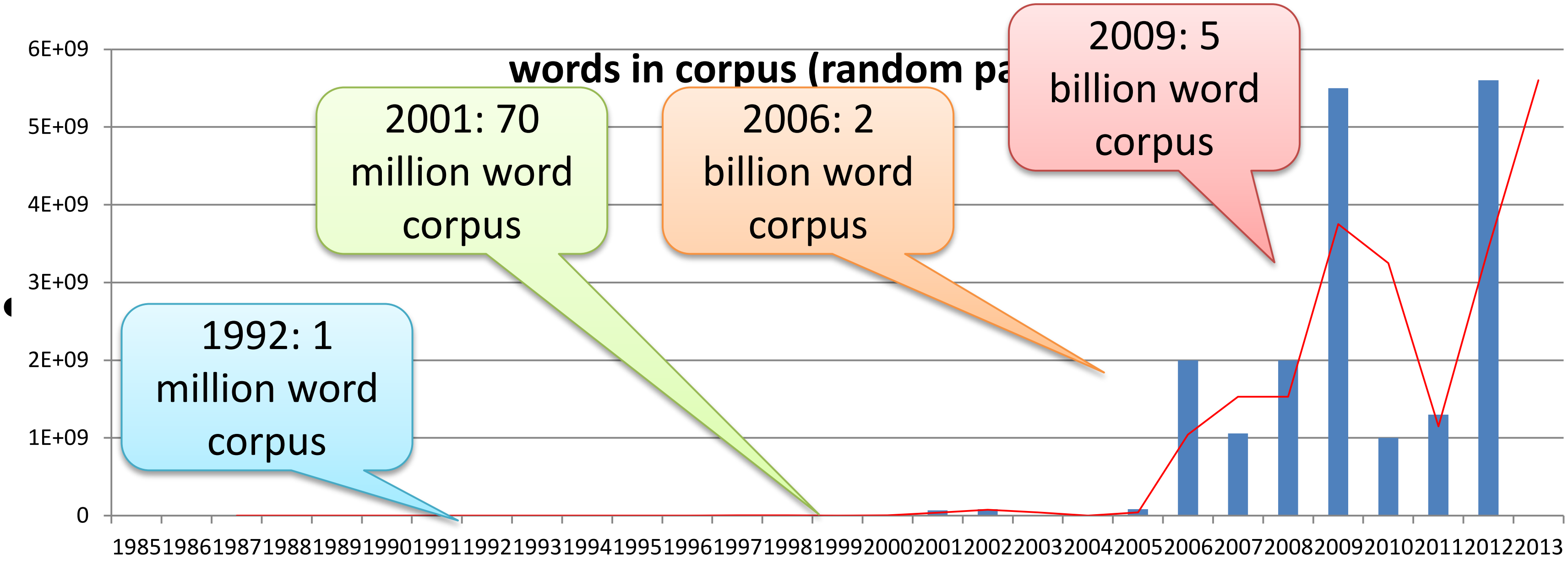




words in corpus (random paper)



papers on large scale corpora

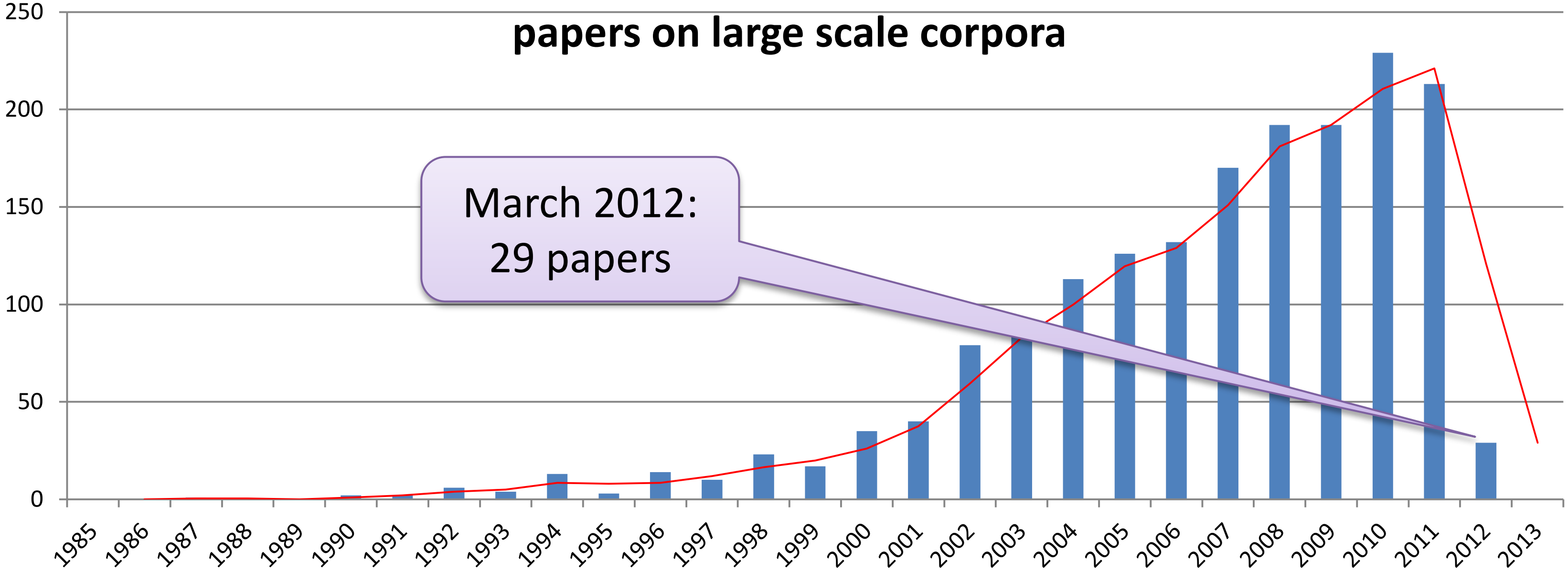


1992: 1 million word corpus

2001: 70 million word corpus

2006: 2 billion word corpus

2009: 5 billion word corpus



March 2012: 29 papers

Do we need large corpora?

- Word frequency decreases in corpus in a semi-quadratic manner:
  - If first most frequent word = 10,000
  - Second frequent word = 4-6,000

  Tendency noticed by George Zipf

  If the corpus is small many words will not be available.

Zipf, George K. 1935. The Psychobiology of Language. Houghton-Mifflin.
Zipf, George K. 1949. Human Behavior and the Principle of Least Effort. Addison-Wesley.

Search Engine Querying

- 2002: Turney and Litman: Sentiment analysis on 100 billion words (estimated part of the Altavista search engine)

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)

Search Engine Querying

- 2002: Turney and Litman: Sentiment analysis on 100 billion words (estimated part of the Altavista search engine)
- Problems:
  - Query per day limit
  - Limited query language (almost not regular expressions)
  - No duplicate filtering

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)

N-gram based corpora

- Google 1T (trillion) 5 gram corpus [1]
- Google Books 155 Billion Word Corpus [2]
- Yahoo! Blog corpus [3] (for Japanese)
  (in development?)

1. Brants, T. and Franz, A. 2006. ``Web 1T 5-gram Version 1", Linguistic Data Consortium, Philadelphia.
2. http://googlebooks.byu.edu/
3. 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.

N-gram based corpora

- Problems
  - Limited context (up to 5 grams, sometimes 7 grams)
  - No additional tagging (POS, dependency structure, NER, etc.)
  - Little usability in linguistic research

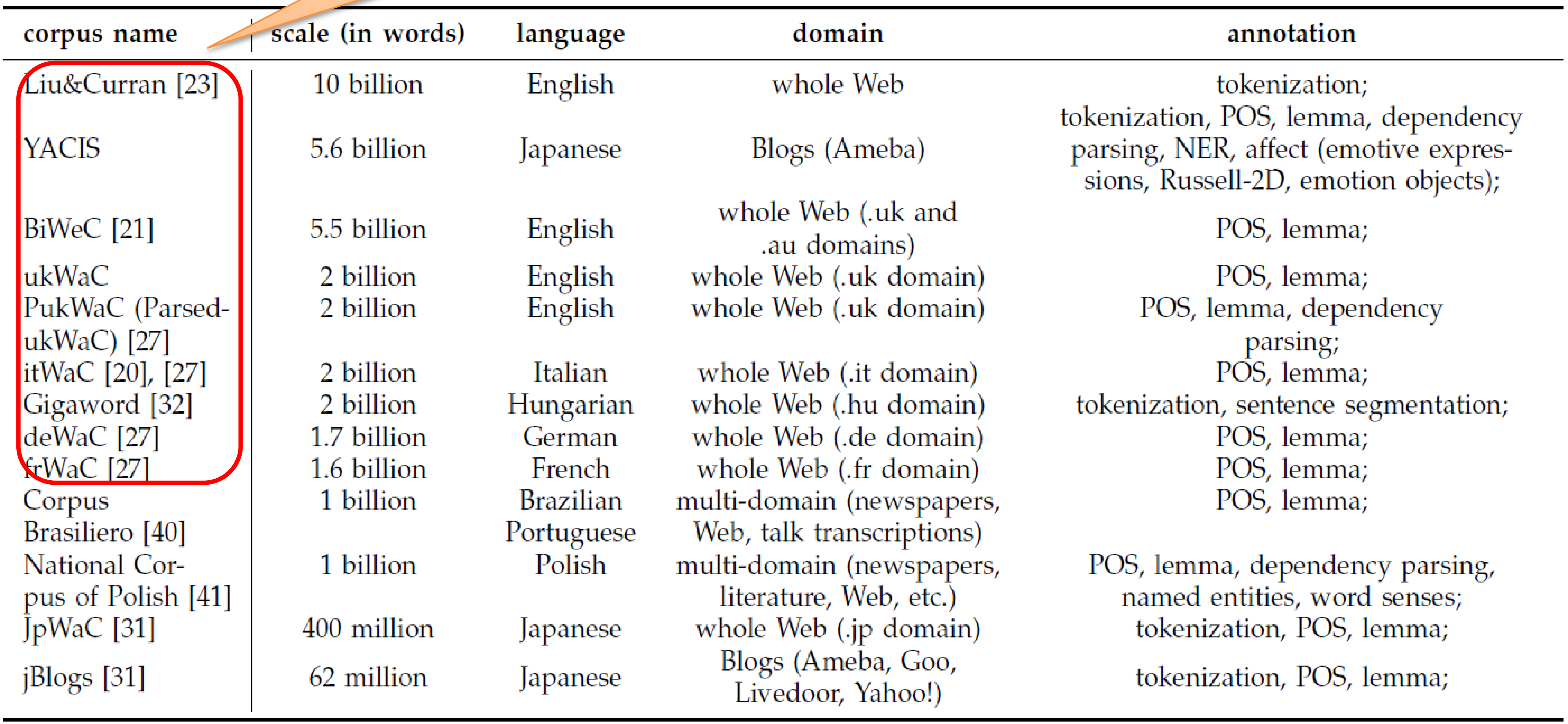
Web-crawled corpora

- Liu&Curran [1] 2006, 10 bil. words, tokenized
- WaCky [2] 2006, 2 bil., POS, lemma, >5 corpora (English, Italian, German, French)
- BiWeC [3], 2009, 5.5 bil., POS, lemma
- YACIS [4], 2010-12, 5.6 bil., POS, lemma, NER, etc.

1. 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.
2. 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.
3. Pomikalek, J., Rychly, 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.
4. Jacek Maciejewski, Michal Ptaszynski, Pawel Dybala, “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), Kitami, Japan/September 2010, pp. 192-195.

Web-crawled corpora

9 / 11 of >1 bil. corpora are Web-crawled

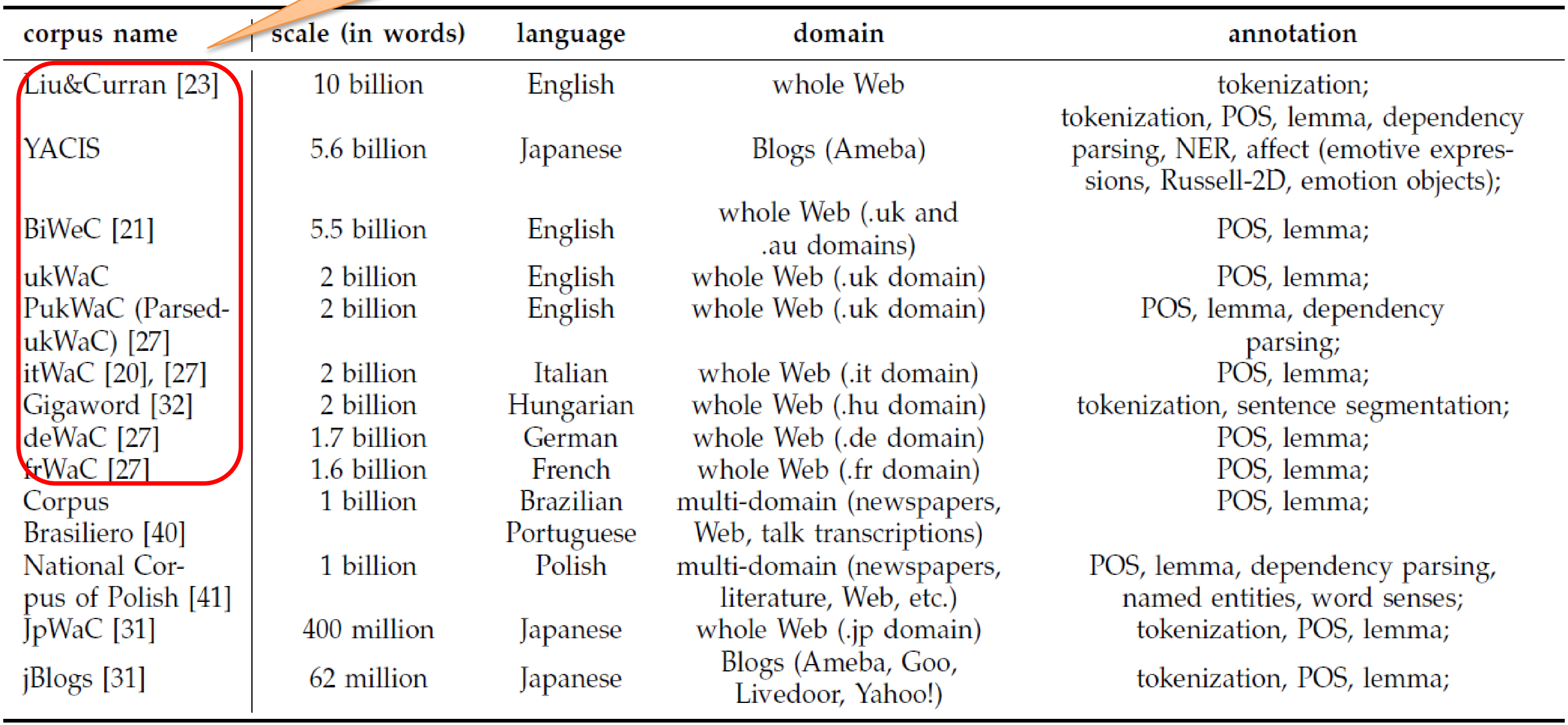


corpus name	scale (in words)	language	domain	annotation
Liu&Curran [23]	10 billion	English	whole Web	tokenization;
YACIS	5.6 billion	Japanese	Blogs (Ameba)	tokenization, POS, lemma, dependency parsing, NER, affect (emotive expressions, Russell-2D, emotion objects);
BiWeC [21]	5.5 billion	English	whole Web (.uk and .au domains)	POS, lemma;
ukWaC	2 billion	English	whole Web (.uk domain)	POS, lemma;
PukWaC (Parsed-ukWaC) [27]	2 billion	English	whole Web (.uk domain)	POS, lemma, dependency parsing;
itWaC [20], [27]	2 billion	Italian	whole Web (.it domain)	POS, lemma;
Gigaword [32]	2 billion	Hungarian	whole Web (.hu domain)	tokenization, sentence segmentation;
deWaC [27]	1.7 billion	German	whole Web (.de domain)	POS, lemma;
frWaC [27]	1.6 billion	French	whole Web (.fr domain)	POS, lemma;
Corpus Brasiliero [40]	1 billion	Brazilian Portuguese	multi-domain (newspapers, Web, talk transcriptions)	POS, lemma;
National Corpus of Polish [41]	1 billion	Polish	multi-domain (newspapers, literature, Web, etc.)	POS, lemma, dependency parsing, named entities, word senses;
JpWaC [31]	400 million	Japanese	whole Web (.jp domain)	tokenization, POS, lemma;
jBlogs [31]	62 million	Japanese	Blogs (Ameba, Goo, Livedoor, Yahoo!)	tokenization, POS, lemma;

Web-crawled corpora

9 / 11 of >1 bil. corpora are Web-crawled

Web presumably contains more text than written data

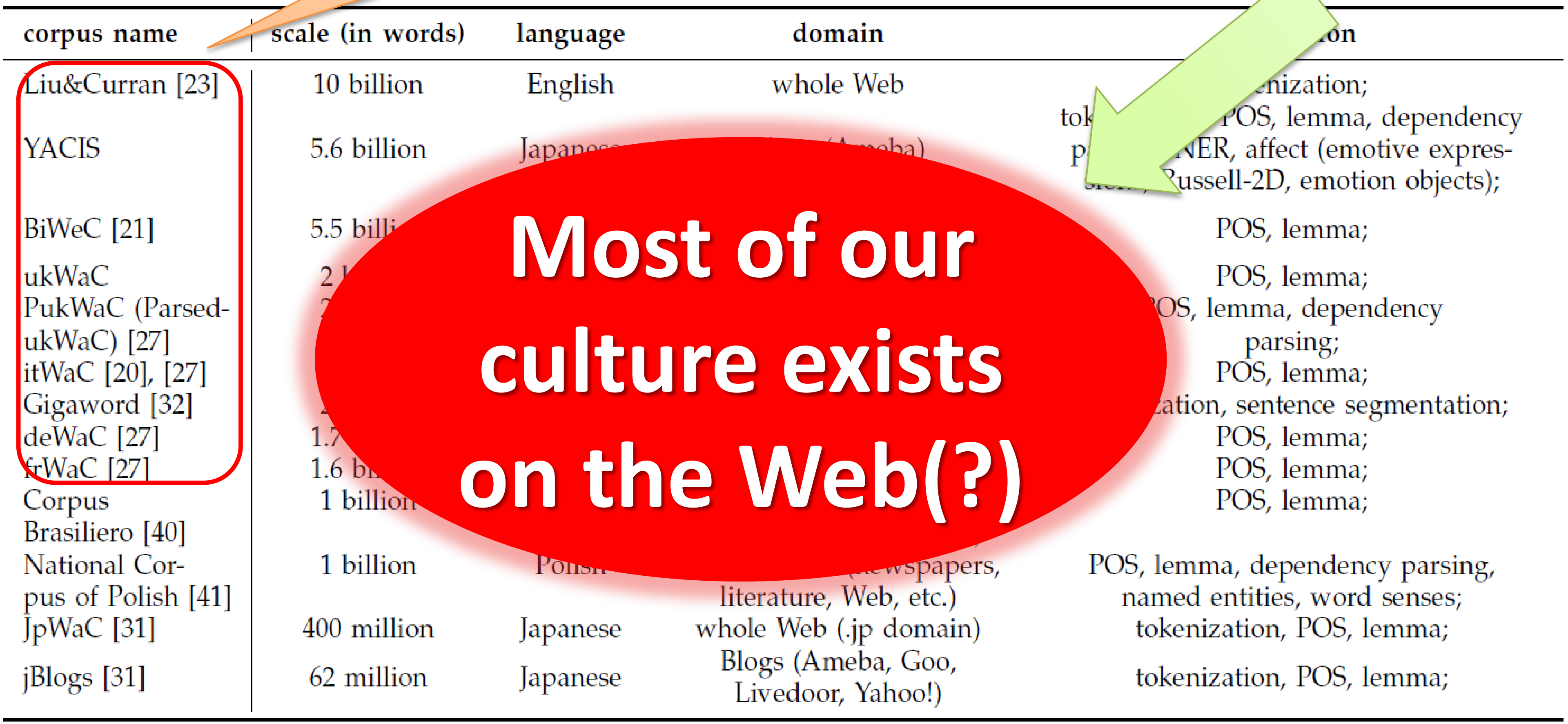


corpus name	scale (in words)	language	domain	annotation
Liu&Curran [23]	10 billion	English	whole Web	tokenization;
YACIS	5.6 billion	Japanese	Blogs (Ameba)	tokenization, POS, lemma, dependency parsing, NER, affect (emotive expressions, Russell-2D, emotion objects);
BiWeC [21]	5.5 billion	English	whole Web (.uk and .au domains)	POS, lemma;
ukWaC	2 billion	English	whole Web (.uk domain)	POS, lemma;
PukWaC (Parsed-ukWaC) [27]	2 billion	English	whole Web (.uk domain)	POS, lemma, dependency parsing;
itWaC [20], [27]	2 billion	Italian	whole Web (.it domain)	POS, lemma;
Gigaword [32]	2 billion	Hungarian	whole Web (.hu domain)	tokenization, sentence segmentation;
deWaC [27]	1.7 billion	German	whole Web (.de domain)	POS, lemma;
frWaC [27]	1.6 billion	French	whole Web (.fr domain)	POS, lemma;
Corpus Brasiliero [40]	1 billion	Brazilian Portuguese	multi-domain (newspapers, Web, talk transcriptions)	POS, lemma;
National Corpus of Polish [41]	1 billion	Polish	multi-domain (newspapers, literature, Web, etc.)	POS, lemma, dependency parsing, named entities, word senses;
JpWaC [31]	400 million	Japanese	whole Web (.jp domain)	tokenization, POS, lemma;
jBlogs [31]	62 million	Japanese	Blogs (Ameba, Goo, Livedoor, Yahoo!)	tokenization, POS, lemma;

Web-crawled corpora

9 / 11 of >1 bil. corpora are Web-crawled

Web presumably contains more text than written data



corpus name	scale (in words)	language	domain	on
Liu&Curran [23]	10 billion	English	whole Web	nization;
YACIS	5.6 billion	Japanes	meba)	tok POS, lemma, dependency p NER, affect (emotive expres- ussell-2D, emotion objects);
BiWeC [21]	5.5 billi			POS, lemma;
ukWaC	2			POS, lemma;
PukWaC (Parsed-ukWaC) [27]	2			OS, lemma, dependency parsing;
itWaC [20], [27]				POS, lemma;
Gigaword [32]	2			ation, sentence segmentation;
deWaC [27]	1.7			POS, lemma;
frWaC [27]	1.6 bi			POS, lemma;
Corpus Brasiliero [40]	1 billion			POS, lemma;
National Corpus of Polish [41]	1 billion	Polish	wspapers, literature, Web, etc.)	POS, lemma, dependency parsing, named entities, word senses;
JpWaC [31]	400 million	Japanese	whole Web (.jp domain)	tokenization, POS, lemma;
jBlogs [31]	62 million	Japanese	Blogs (Ameba, Goo, Livedoor, Yahoo!)	tokenization, POS, lemma;

Most of our culture exists on the Web(?)

Japanese Web/blog-based corpora

- YACIS
- JpWaC
- jBlogs
- KNP
- Kawahara&Kurohashi
- Yahoo! Blog corpus

1. 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.
2. 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.
3. Chikara Hashimoto, Sadao Kurohashi, Daisuke Kawahara,Keiji Shinzato and Masaaki Nagata, "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, **2011**.
4. 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.
5. 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.

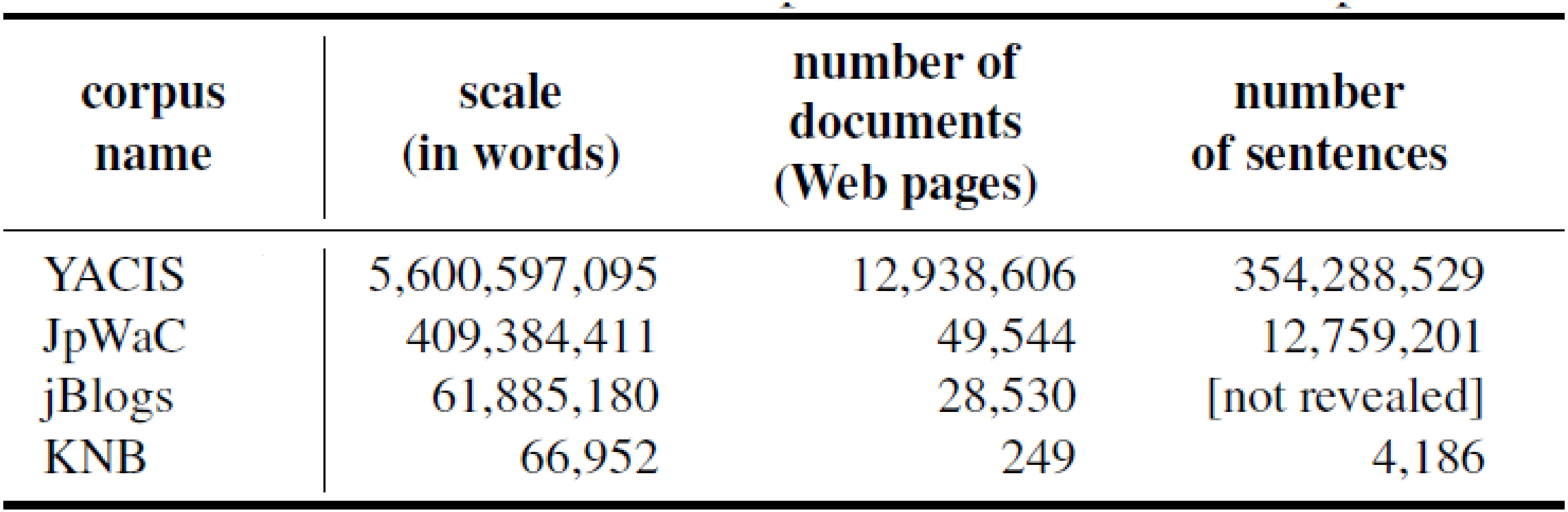
Japanese Web/blog-based corpora

- YACIS
- JpWaC [1]
- jBlogs [2]
- KNP [3]
- Kawahara&Kurohashi [4]
- Yahoo! Blog corpus [5]

Could not find detailed information on these

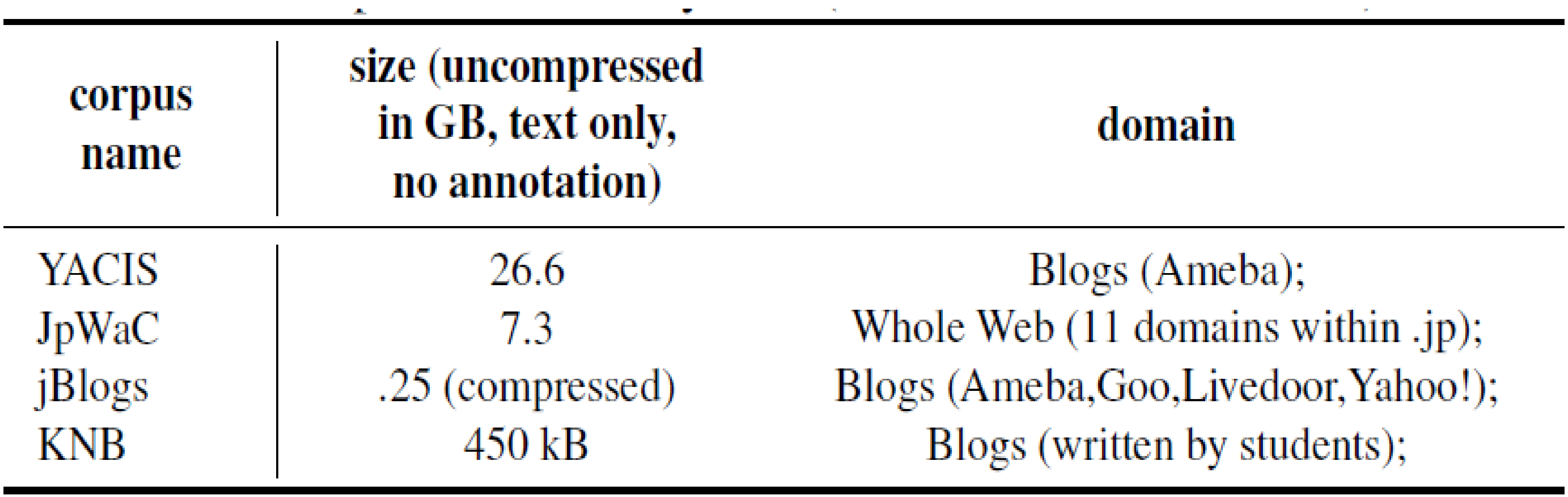
1. Erjavec, I. S., Erjavec, T., Kilgarriff, A. 2008. ``A web corpus and word sketches for Japanese", Information and Media Technologies, 3(3), pp. 529-551.
2. Baroni, M. and Ueyama, M. 2006. ``Building General- and Special-Purpose Corpora by Web Crawling", In Proceedings of the 13th NIJL International Symposium on Language Corpora: Their Compilation and Application.
3. Chikara Hashimoto, Sadao Kurohashi, Daisuke Kawahara,Keiji Shinzato and Masaaki Nagata, "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, **2011**.
4. 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.
5. 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.

Japanese Web/blog-based corpora



corpus name	scale (in words)	number of documents (Web pages)	number of sentences
YACIS	5,600,597,095	12,938,606	354,288,529
JpWaC	409,384,411	49,544	12,759,201
jBlogs	61,885,180	28,530	[not revealed]
KNB	66,952	249	4,186

Japanese Web/blog-based corpora



corpus name	size (uncompressed in GB, text only, no annotation)	domain
YACIS	26.6	Blogs (Ameba);
JpWaC	7.3	Whole Web (11 domains within .jp);
jBlogs	.25 (compressed)	Blogs (Ameba,Goo,Livedoor,Yahoo!);
KNB	450 kB	Blogs (written by students);

Conclusions

- Showed statistics of papers on large corpora and size of corpora
  - Number of papers increases linearly
  - Size of corpora increases suddenly
- If corpus size is small many words will be not appear at all (Zipf, 1935)

Conclusions

- Presented survey on research on large scale corpora based on:
  - Search Engine Querying
  - N-gram based corpora
  - Web-crawled corpora
  - Japanese Web-based corpora
- A few >1 bil. corpora
- Usual annotations: POS, lemma

Future Work (in general)

- Set an up-to-date standard for corpora
  - > 2 bil. Words (?)
- Annotate with all available information
- Apply!

Thank you for your attention!

Michal Ptaszynski

ptaszynski@ieee.org