Information and Media Technologies 9(4): 429–445 (2014)
reprinted from: Computer Software 31(2): 151–167 (2014)

A Survey on Large Scale Corpora and Emotion Corpora

Michal Ptaszynski, Rafal Rzepka, Satoshi Oyama, Masahito Kurihara and Kenji Araki

In this paper we present a survey on natural language corpora, with particular focus on corpora of large scale and those applicable to sentiment analysis. Natural language corpora are crucial for training various Software Engineering applications, from part-of-speech taggers and dependency parsers to dialog systems or sentiment analysis software. We compare several natural language corpora created for different languages, analyze their distinctive features and the amount of additional annotations provided by the developers of those corpora.

1 Introduction

Recent need for handling Big Data in Artificial Intelligence (AI) applications has introduced a new set of challenges for Software Engineering (SE). One of them is efficient and fast access to rich in information large scale resources, such as annotated natural language corpora. Natural language corpora are crucial for training many AI applications, from part-of-speech taggers and dependency parsers to dialog systems or sentiment analysis software. Often natural language corpora used in such applications are of small scale. However, recently there have been several initiatives to create locally accessible large scale corpora, which could become useful for various applications. In this paper we present a survey on such natural language corpora, with particular focus on corpora of large scale and those applicable to sentiment analysis. We compare a number of such corpora created for different languages, analyze their distinctive features and the amount of additional annotations provided by the developers of those corpora. Additionally, we pay a special attention to such corpora created for the Japanese language, since our research is in large part performed for this language.

The outline of this paper is as follows. In Section 2 we discuss natural language corpora in general, their importance in research and present need for large scale corpora. In section 3 we describe in detail the scope of this survey.Section 4 presents the main part of this paper, namely, survey in large scale corpora and emotion corpora, separately. To make the survey more thorough and robust, in Section 5 we present a more detailed comparison of several corpora described in previous section. We compare three large scale corpora of different languages, but comparable sizes, and three other corpora of the same language but different sizes (large medium and small). In Section 6 we present a discussion on availability of corpora. Finally, in Section 7 we summarize and conclude the survey.

2 Natural Language Corpora

Text corpora are some of the most vital linguistic resources in Artificial Intelligence (AI) and Natural Language Processing (NLP). These include newspaper corpora, like Wall Street Journal Corpus [1] or Mainichi Shinbun Corpus [2], conversation corpora, like the BC3 corpus [3] or the CSJ corpus

大規模コーパス及び感情コーパスに関する調査.

Michal Ptaszynski, 北見工業大学情報システム工学科, Department of Computer Science,Kitami Institute of Technology.

Rafal Rzepka, 小山　聡, 栗原正仁, 荒木健治, 北海道大学大学院情報科学研究科, Graduate School of Information Science and Technology,Hokkaido University.

コンピュータソフトウェア, Vol.31, No.2 (2014), pp.151–167.
[解説論文] 2013 年 4 月 22 日受付.

[4], as well as corpora of literature, such as Aozora Bunko [5]. The importance of corpora is widely recognized and numerous corpora have been compiled so far for different languages. However, comparing to major world languages, like English, there are few large corpora available for the Japanese language [6]. Moreover, grand majority of them is usually unsuitable for sentiment analysis. Although there exist speech corpora, such as Corpus of Spontaneous Japanese [4], which could become suitable for emotion processing research, due to the difficulties with compilation of such corpora they are relatively small.

Switching to 64bit-based machines and environments has allowed compilation and efficient processing of billion-word and larger corpora. Recently the corpora based on automatically extracted Web contents gain on popularity. This is due to the potential of such Web-based corpora to solve the problem of corpus small size. There exist several initiatives in the creation of large-scale corpora, such as WaCky[†1], and the need for those is constantly growing. For the Japanese language there also exist somewhat large Web-based corpora (containing several million words), such as JpWaC [6] or jBlogs [7]. However, access to them is usually allowed only from Web interface, and they are not fully annotated. Finally, although there exist very large resources, like Google N-gram Corpus [8], the textual data sets in such resources are usually short (up to 7-grams) and do not contain any contextual information (such as "snippets" with the closest content a particular n-gram appears in). This makes them unsuitable for sentiment analysis and emotion-related research, since most of contextual information, so important in expressing emotions and opinions, is lost.

In this paper we present the above mentioned initiatives. In the following sections we firstly describe some of the corpora, compare their scale, features and the amount of additional annotations performed on them. We also dedicate a separate section to compare in more detail corpora of similar features.

3 Scope of this Survey

In this paper we focused on two different types of corpora, large scale corpora and emotion corpora. The survey of each of this type needed a separate set of restrictions to be assumed before performing the survey.

Firstly, for large scale corpora we assumed that "large scale" means at least one billion words or more. An exception to this assumption was made in the paragraph describing Japanese language corpora, which are usually of small or medium scale. Another restriction was that the corpus needed to be either 1) widely available or accessible, for example through a Web API, etc., or 2) thoroughly described, so there was enough data for a detailed comparison with other corpora. In practice this limited the scope of this part of the survey mainly to Web-based corpora, with some exceptions for corpora of mixed domains (for example Corpus Brasiliero [9] was gathered from newspapers, Web and talk transcriptions, etc.).

In the section dedicated to emotion corpora we made less assumptions due to the small number of such corpora. The only assumption was that we defined "corpus" as "collection of sentences". Language resources that are not corpora *per se*, such as ontologies like WordNet[†2], or lexicions, are also sometimes called corpora. This is acceptable when a general definition of "corpus" is applied ("collection of text"). The only exception we made was for Emotive Expression Database, which is available as a lexicon, but was based on a collection of sentences (an emotive expression dictionary).

4 Corpora Survey

This section presents the survey of some of the most relevant recent research on corpora. We divided the description of the research into "Large-Scale Corpora" (related to natural language corpora in general) and "Emotion Corpora" (related to sentiment analysis research).

4.1 Large-Scale Corpora

The notion of a "large scale corpus" has appeared in linguistic and computational linguistic literature

[†1] http://wacky.sslmit.unibo.it/

[†2] http://wordnet.princeton.edu/

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 recent corpora. For example, Sasaki et al. [10] 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 [11], one of the copora described here consists of 12,938,606 documents (blog articles). The rough estimation indicates that the corpus of Sasaki et al. covers less than 5% of YACIS (in particular 4.08%). Therefore we mostly focused on research scaling the meaning of "large" up to around billion-words and more.

Firstly, we need to address the question of 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 [12] 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 and developed by George Kingsley Zipf in late 1930's to 1940's [13][14] and 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 large enough, many words will not appear in it at all. Baroni and Ueyama [7], and Pomikálek et al. [15] 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 language phenomena and provide statistically reliable proofs for those phenomena. There are opponents of uncontrolled over-scaling of corpora, such as Curran (with Osborne in [16]), 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 scale of a corpus is an important feature in corpus linguistic research and eventually join the initiative of developing a 10-billion word corpus of English (see Liu and Curran [17]).

Liu and Curran [17], followed by Baroni and Ueyama [7], indicate at least two types of research dealing with large-scale corpora. First is using popular search engines, such as Google[†3] or Yahoo![†4]. Second is crawling the World Wide Web and downloading its contents for further analysis. The latter comes also in two kinds: N-gram based corpora with limited word context (up to pentagram, etc.), and copora with local access to full context (document, sentence, etc.). In the sections below we will describe the most representative corpora of each of those kinds.

4.1.1 Search Engine Querying

In this type of research one considers as "corpus" the contents of the World Wide Web and uses popular search engines to search for certain keywords (words) or keywords sequences (phrases). Search engine APIs usually present at least two kinds of information: 1) estimates of hit counts for keywords, and 2) wider contexts of keywords, called "snippets" (a short, about three-line-long text containing the keyword). The first kind of information allows for statistical analysis of the searched keywords. The second one allows further analysis of the snippet contents, such as calculating relevance between searched keywords and other contents. Research applying this kind of information retrieved from the Web is generally referred to as the "Web mining" field. There has been a wide range of applications for Web mining, beginning with text classification [18], or predicting word relationship [19], to the most recent ones, such as affect analysis [20], or detecting cyberbullying entries [21]. One example, in which the authors directly referred to the Web contents as "corpus", is the research by Turney and Littman [22]. 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 [†5].

Unfortunately, this kind of research is inevitably constrained with limitations of the search engine's

[†3] http://www.google.com

[†4] http://www.yahoo.com

[†5] In 2004 AltaVista (http://www.altavista.com/) became a part of Yahoo!.

API. Pomikálek et al. [15] indicate a long list of such limitations. Some of them include: limited query language (e.g. no search by sophisticated 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 for 24 hours - an unacceptable limitation in linguistic research), search queries are ordered with a manner irrelevant to linguistic research, etc. Kilgariff [23] 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, a great disadvantage of all research based on Web mining is the fact that the results of such research are not reproducible, since the information in search engines is frequently updated. Finally, 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 World Wide Web, and argues that this is the optimal way of utilizing the Internet contents for research in linguistics and computational linguistics.

4.1.2 N-gram Based Corpora

One of the attempts to deal with the problem of irreproducibility of research results in the Web mining field was the appearance of n-gram based corpora created by crawling the World Wide Web. Two large scale corpora of this kind have been presented by Google. One is the "Web 1T (trillion) 5 gram" English corpus [24], published in 2006. It is estimated to contain one trillion of tokens and 95 billion sentences gathered from the Web. 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 context. The second one is the "Google Books 155 Billion Word Corpus" [†6] published in 2011. It contains 1.3 million books in English 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[†7], unfortunately only in the n-gram form (no context larger than 5-gram). Four years after the above Google 1T corpus, Microsoft presented a similar Web N-gram Corpus [25], containing 1.4 trillion words. Both, Google 1T and Microsoft N-gram corpora have been collected from the contents of Web pages indexed by either Google Search Engine or Microsoft's Bing Search Eengine.

The potential range of applications of n-gram based large scale corpora is similar to general Web mining, with a restriction to narrow context (no sentences or snippets are available, only n-grams). In practice, most popular applications include such tasks as word segmentation [25], word sense disambiguation [26], or spelling correction [27].

4.1.3 Web-crawled Corpora

Among corpora created with Web crawling methods, Liu and Curran [17] 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. The corpus was created using two techniques: IP Address Sampling (a technique randomly generating IP addresses) and Random Walk (a technique which produces an approximated undirected web graph with uniform samples of nodes (Web pages)) using as seed words contents of the Open Directory[†8]. This corpus 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) [7] [28] project started gathering and linguistically processing large scale corpora from the Web. In the years 2005-2007 the project resulted in more then five collections of around two billion word corpora for different languages, such as English (ukWaC), French (frWaC), German (deWaC) or Italian (itWaC). The corpora have been created using about thousand random words of medium occurrence frequency for each language as seed words. The seed words were then queried in a search engine to obtain seed URLs. The seed URLs pro-

[†6] http://googlebooks.byu.edu/

[†7] http://books.google.com/ngrams/datasets

[†8] http://www.dmoz.org

Table 1 Comparison of different corpora, ordered arbitrarily by size (number of words/tokens).

corpus name	scale (in words)	language	domain	annotation
Liu&Curran [17]	10 billion	English	whole Web	tokenization;
YACIS [11]	5.6 billion	Japanese	Blogs (Ameba)	tokenization, POS, lemma, dependency parsing, NER, affect (emotive expressions, valence, activation, emotion objects);
BiWeC [15]	5.5 billion	English	whole Web (.uk and .au domains)	POS, lemma;
ukWaC [28]	2 billion	English	whole Web (.uk domain)	POS, lemma;
PukWaC (Parsed-ukWaC) [28]	2 billion	English	whole Web (.uk domain)	POS, lemma, dependency parsing;
itWaC [7][28]	2 billion	Italian	whole Web (.it domain)	POS, lemma;
Gigaword [29]	2 billion	Hungarian	whole Web (.hu domain)	tokenization, sentence segmentation;
deWaC [28]	1.7 billion	German	whole Web (.de domain)	POS, lemma;
frWaC [28]	1.6 billion	French	whole Web (.fr domain)	POS, lemma;
Corpus Brasiliero [9]	1 billion	Brazilian Portuguese	multi-domain (newspapers, Web, talk transcriptions)	POS, lemma;
National Corpus of Polish [30]	1 billion	Polish	multi-domain (newspapers, literature, Web, etc.)	POS, lemma, dependency parsing, named entities, word senses;
JpWaC [6]	400 million	Japanese	whole Web (.jp domain)	tokenization, POS, lemma;
jBlogs [6]	62 million	Japanese	Blogs (Ameba, Goo, Livedoor, Yahoo!)	tokenization, POS, lemma;

Table 2 Comparison of different Japanese corpora, ordered by the 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	5,600,597,095	12,938,606	354,288,529	26.6	Blogs (Ameba);
JpWaC [6]	409,384,411	49,544	12,759,201	7.3	whole Web (11 different domains within .jp);
jBlogs [7]	61,885,180	28,530	[not revealed]	.25 (compressed)	Blogs (Ameba, Goo, Livedoor, Yahoo!);
KNB [45]	66,952	249	4,186	450 kB	Blogs (written by students exclusively for the purpose of the research);
Minato et al. [51]	14,202	1	1,191	[not revealed]	Dictionary examples (written by dictionary author);

vide initial contents (Web pages) for the corpus and links to next URLs. Retrieved Web pages are cleaned from HTML code, and duplicate pages are deleted. The contents are then tokenized and annotated with parts of speech. The corpora have been successfully applied in tasks such as collocation extraction or translation pair extraction. 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 [15], or **Big Web Corpus** has been collected from the whole Web contents in 2009 and consists of about 5.5 billion words in English. The authors of this corpus aimed to go beyond the border of 2 billion words set by the WaCky initiative[†9] as a borderline for corpus processing feasibility for modern (32-bit) software. Most tools and precedures for creating BiWeC were borrowed from the WaCky project. Despite its scale and great potential, no applications of the corpus have been re-

[†9] http://wacky.sslmit.unibo.it/

ported so far.

Billion-word scale corpora have been recently developed also for other, less resourced languages, such as Brazilian Portuguese [9], Hungarian [29] or Polish [30]. All corpora described in this section are compared in Table 1.

4. 1. 4 Japanese Web-crawled Corpora

As for large corpora in Japanese, despite the fact that Japanese is a well recognized and described world language, there have been only few corpora of a reasonable size. Below we describe those corpora in detail. Some comparable information about the corpora is also represented in Table 2.

YACIS or **Y**et **A**nother **C**orpus of **I**nternet **S**entences was collected automatically by Maciejewski, Ptaszynski and Dybala [31] from the pages of Ameba blog service, and developed further by Ptaszynski et al. [11] [32]. It contains 5.6 billion words within 350 million sentences. The compilation process was performed within 3 weeks between 3rd and 24th of December 2009. Maciejewski et al. extracted only pages containing Japanese posts (pages with legal disclaimers or written in languages other than Japanese were omitted). In the initial phase they provided their crawler, optimized to crawl only Ameba blog service, with thousand links taken from Google (response to single query: 'site:ameblo.jp'). They saved all pages to disk as raw HTML files (each page in a separate file) and afterward extracted all the posts and comments from HTML, and divided them into sentences. The original structure (blog post and comments) is preserved, thanks to which semantic relations between posts and comments are retained. Further annotation by Ptaszynski et al. [11] [32] has been performed within the years 2010-2012, and contains annotations of tokens, lemmas, parts-of-speech, dependency structures and named entities [11]. YACIS has been applied in a number of research, mostly emotion related. Some of them include emotional and ethical information retrieval [33] [34] or automatic generation of emotion object database [35]. The corpus has also been annotated with emotion-related information by Ptaszynski et al. [32] and is discussed in the next section.

Apart from YACIS, Srdanović Erjavec et al. [6] used WaC (Web as Corpus) Toolkit[†10] and Kilgariff et al.'s Sketch Engine [36], a tool for thesauri generation from large scale corpora. They gathered **JpWaC**, a 400 million word corpus of Japanese. Although JpWac covers only about 7% of YACIS (400 mil. words vs 5.6 bil. words), it shows that freely available tools developed for European languages are to some extend applicable also for languages of completely different typography, like Japanese[†11]. However, the researchers faced several problems, such as normalization of character encoding for all web pages[†12]. For comparison, YACIS is based on Ameba blog service which is encoded by default in Unicode. The corpus has been applied mostly in language learning for collocation generation and sentence example extraction.

Baroni and Ueyama [7] 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 jBlogs, Baroni and Ueyama noticed that blog posts contained many Japanese emoticons, or *kaomoji*[†13]. They report 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 parts of speech distribution, 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. For comparison, in YACIS the researchers dealt with this problem by using CAO, a system for detailed analysis of Japanese emoticons developed previously by Ptaszynski et al. [37]. Except detailed morphological analysis of the corpus, no practical applications of jBlogs have been reported so far.

[†10] http://www.drni.de/wac-tk/

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

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

[†13] For more detailed description of Japanese emoticons, see [37].

Table 3 Comparison of emotion corpora ordered by the amount of annotations (abbreviations: T=tokenization, POS=part-of-speech tagging, L=lemmatization, DP=dependency parsing, NER=Named Entity Recognition).

corpus	scale	language	annotated affective information						syntactic
name	(in sentences / docs)		emotion classes (standard)	emotive expressions	emotive/ non-emot.	valence/ activation	emotion intensity	emotion objects	annotations
YACIS [32]	354 mil. /13 mil.	Japanese	10 (language and culture based)	○	○	○/○	○	○	T,POS,L, DP,NER;
Ren-CECps1.0 [42]	12,724 /500	Chinese	8 (Yahoo! news annotation standard)	○	○	○/×	○	○	T,POS;
MPQA [43]	10,657 /535	English	none (no standard)	○	○	○/×	○	○	T,POS;
KNB [45]	4186 /249	Japanese	none (no standard)	○	×	○/×	×	○	T,POS,L, DP,NER;
Minato et al. [51]	1,190 separate sentences	Japanese &English	8 (chosen subjectively)	○	○	×/×	×	×	POS;
Aman&Szpakowicz [40]	5205 /173	English	6 (face recognition)	○	○	×/×	○	×	×
Das&Bandyopadhayay [46]	12,149 /123	Bengali	6 (face recognition)	○	×	×/×	○	×	×
WKEC [38][47]	20,500 separate sentences	Japanese	4 (Fischer's emotion systematic tree)	○	×	×/×	×	×	T,POS;
Mishne [50]	815,494 blog posts	English	132 (LiveJournal annotation standard)	×	×	×/×	×	×	×

4.2 Emotion Corpora

The research on sentiment analysis and opinion mining, as well as related research on affect analysis, or affect sensing from text, has resulted in a number of sentiment and affect analysis systems developed within several years [38][39][40][41]. Unfortunately, most of such research usually ends in proposing and evaluating a certain system. The real world application that would be desirable, such as annotating affective information on linguistic data is limited to processing a usually small test sample in the evaluation of the system. The small number of annotated emotion corpora that exist are mostly of limited scale and are annotated manually. Below we describe and compare some of the most notable emotion corpora. As an interesting remark, eight out of nine emotion corpora described below are extracted from blogs. The corpora are also compared in Table 3.

YACIS [31] has been described shortly above in section 4.1.4. Originally it was not created as an emotion corpus, but as a general blog corpus. Ptaszynski et al. [32] performed on YACIS further annotation of emotion labels on sentences, emotive expressions, emotion valence and emotion intensity. The annotations were performed automatically with the use of an affect analysis system ML-Ask [39]. Later, Ptaszynski et al. [35] added annotations of emotion objects. The emotion objects were extracted automatically by using an emotional consequence retrieval method designed originally as a search engine based Web mining technique [20]. Originally the technique was extracting emotional associations for an input phrase or sentence (considered as the cause or object of the emotion). Ptaszynski et al. [35] reversed the method to extract emotion objects for emotional expressions. The emotion annotated version of YACIS is available together with the original corpus. The emotion object database is also available as a separate sub-corpus of YACIS.

Quan and Ren [42] created a Chinese emotion blog corpus called **Ren-CECps1.0**†14. They collected 500 blog articles from various Chinese blog services, such as sciencenet blog†15 or qq blog†16. The articles were manually annotated with a large variety of information, such as emotion class, emotive expressions, polarity level, or emotion object. Although the syntactic annotations were simplified to tokenization and POS tagging, this corpus, next to YACIS in Japanese, is one of the most representative when it comes to the overall variety of annotations.

Wiebe et al. [43] [44] report on creating the **MPQA** corpus of news articles. The corpus contains 10,657 sentences in 535 documents†17 extracted from a wide variety of news sources from June 2001 to May 2002 and annotated manually by trained annotators. The annotation schema includes a variety of emotion-related information, such as emotive expressions, emotion valence, intensity, etc. However, Wiebe et al. focused on detecting subjective (emotive) sentences and classifying them into positive and negative. Thus their annotation schema, although one of the richest, does not include emotion classes (joy, fear, anger, etc.).

A corpus of Japanese blogs with large amount of annotated information has been developed by Hashimoto et al. in 2010 and published in 2011 [45]. The corpus was developed jointly by the National Institute of Information and Communications Technology, Kyoto University, and the NTT Communication Science Laboratories. The **KNB**†18 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 proposed a certain standard for preparing corpora, especially blog corpora for sentiment and affect-related studies. The corpus contains all relevant syntactic and morphological annotations, including POS tagging, dependency parsing or named entities. 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 could select their words more carefully than they would in private.
Aman and Szpakowicz [40] also constructed a small-scale English blog emotion corpus. However, they focused not on the grammatical annotations, but on the affect-related annotations. The corpus was created in the following way. Firstly, seed words were selected for six emotion classes borrowed from Ekman's standard for basic emotions in facial expression recognition. Using the seed words, Aman and Szpakowicz automatically retrieved blog posts containing the seed words. Then the blog posts were annotated manually by two layperson annotators for each blog post. As an interesting remark, Aman and Szpakowicz were some of the first to recognize the task of distinguishing between emotive and non-emotive sentences. This problem is usually one of the most difficult in text-based affect analysis and is therefore often omitted in such research.

Das and Bandyopadhyay constructed an emotion annotated corpus of blogs in Bengali [46]. The corpus contains 12,149 sentences within 123 blog posts extracted from Bengali web blog archive†19. Similarly to Aman and Szpakowicz, Das and Bandyopadhyay also annotated their corpus with Ekman's six class emotion annotation standard for face recognition.

Matsumoto et al. [38] [47] report on construction of *Wakamono Kotoba* (Slang of the Youth) Emotion Corpus (or **WKEC**) for the Japanese language. The corpus contains separate sentences extracted manually from *Yahoo! blogs*†20. Each sentence contains at least one word from a slang lexicon and one word from an emotion lexicon, with additional emotion class tags added per sentence.

†14 Abbreviation of **Ren**'s **C**hinese blogs **E**motion **C**orpus.

†15 http://blog.sciencenet.cn/

†16 http://blog.qq.com/

†17 The new MPQA Opinion Corpus version 2.0 contains additional 157 documents, 692 documents in total.

†18 Abbreviation of **K**yoto University and **N**TT Lab **B**log Corpus.

†19 http://www.amarblog.com/

†20 http://blog.search.yahoo.co.jp/

Both lexicons were created manually. Originally [38], the emotion class set used for annotation (9 emotion classes) was set by applying the 6 emotion types from face recognition research [48] and adding additional three of their subjective choice. Later [47], Matsumoto et al. enlarged their corpus and refined the emotion class set. They used four emotion classes, basing on Fischer's emotion systematic tree [49], which were a rough generalization of the original class set.

Mishne [50] collected a corpus of English blogs from the LiveJournal[†21] blog service. The corpus contains 815,494 blog posts, from which many are annotated with emotions (moods) by the blog authors themselves. The LiveJournal service offers an option for its users to annotate their mood while writing the blog. The list of 132 moods includes words like "amused", or "angry". The LiveJournal mood annotation standard offers a rich vocabulary to describe the writer's mood. However, this richness, without any generalization to emotion classes, has been considered too troublesome, both for users in making the annotation choice, and for researchers to generalize the data in a meaningful manner [42].

Minato et al. [51] [52] collected a 14,202 word- / 1,190 sentence-corpus and annotated it manually. The corpus is a collection of dictionary examples from "A short dictionary of feelings and emotions in English and Japanese" [53]. It is a dictionary created for the need of Japanese language learners. Unfortunately, the dictionary does not propose any coherent emotion class list, but rather the emotion concepts are chosen subjectively. Although the corpus by Minato is the smallest corpus of all mentioned in this section, differently to others Minato et al. provide a full statistical analysis of the corpus. Therefore in this paper we use their research as one of the Japanese emotion corpora for detailed comparison. The detailed comparison is given in section 5. 2.

Finally, Ptaszynski et al. [20] [39] gathered an **Emotive Expression Database** (**EED**). It is not a corpus *per se*, but was converted into an emotive expression database by Ptaszynski et al. in their research on affect analysis of utterances in Japanese. They based the corpus on "Emotive Expression Dictionary" [54]. The original dictionary was developed by Akira Nakamura in a period of over 20-year time. It is a collection of over two thousand expressions describing emotional states collected manually from a wide range of literature. Nakamura's dictionary also proposes a 10-type classification of emotions, which reflect the Japanese language and culture. This classification is applied to the lexicon itself. All expressions are classified as representing a specific emotion type, one or more if applicable. The Emotive Expression Database was also one of the lexical resources used in annotation of sentiment labels on YACIS, mentioned above.

5 Detailed Comparison of Corpora

In previous sections we performed a general survey on natural language corpora with an overview and comparison of general features of the corpora. In this section we perform a further detailed analysis and comparison. We compared those corpora for which either detailed descriptions existed or the corpora were freely available for download and analysis. The comparison is performed separately for large scale corpora and emotion corpora.

5. 1 Comparison of Large Scale Corpora

To obtain a wider view on the corpora we compared large scale corpora in two ways. Firstly we compared part-of-speech (POS) tagging on corpora of comparable size (large scale corpora of 1 bil. tokes and more) but different languages to analyze differences of POS distribution between languages. In particular we compared Japanese language to British and Italian. Secondly, we compared POS distributions for corpora of the same language (Japanese), but of different sizes (small, medium, large).

In particular we compared YACIS (large), jBlogs (medium) and JENAAD (small size newspaper corpus). Information on the distribution of parts of speech is represented in Table 4.

Japanese vs British and Italian:

The comparison of three large scale corpora (YACIS [31] in Japanese, ukWaC [28] in British English and itWaC in Italian) revealed interesting observations. Although not all information on part-of-speech statistics is provided for the latter two corpora, the available information shows interest-

[†21] http://www.livejournal.com/

Table 4 Comparison of POS distribution across corpora.

Part of	YACIS		jBlogs	JENAAD	ukWaC	itWaC
speech	percentage	number	percentage	percentage	number	number
Noun	34.69%	(1,942,930,102)	34%	43%	1,528,839	941,990
Particle	23.31%	(1,305,329,099)	18%	26%	[not provided]	[not provided]
Verb	11.57%	(647,981,102)	9%	11%	182,610	679,758
Auxiliary verb	9.77%	(547,166,965)	7%	5%	[not provided]	[not provided]
Adjective	2.07%	(116,069,592)	2%	1%	538,664	706,330
Interjection	0.56%	(31,115,929)	<1%	<1%	[not provided]	[not provided]
Other	18.03%	(1,010,004,306)	29%	14%	[not provided]	[not provided]

ing differences between part-of-speech distribution among languages†22. In all compared corpora the largest is the number of "nouns". However, differently to all Japanese corpora, second frequent part of speech in British English and Italian corpus was "adjective", while in Japanese it was "verb" (excluding "particles"). This difference is the most vivid in ukWaC. The differences could be influenced by the phenomenon of ambiguities in classification of Japanese adjectives, which has been noticed in linguistic literature for several years (see for example Backhouse in [58]). In the discussion it is argued that a large group of Japanese adjectives (such as *i*-adjectives, etc.) behave grammatically in a similar way to Japanese verbs (verb-like conjugation). Computer analysis of this phenomenon could reveal more details on this phenomenon and contribute to the fields of language anthropology, and the philosophy of language in general. **YACIS vs jBlogs and JENAAD:**

We compared tendencies in POS annotations between YACIS, jBlogs (both mentioned in section 4.1) and JENAAD [55]. The latter is a medium-scale corpus of newspaper articles gathered from the Yomiuri daily newspaper (years 1989-2001). It contains about 4.7 million words (approximately 7% of jBlogs and 0.08% of YACIS). The comparison of those corpora provided interesting observations. Parts of speech in jBlogs and JENAAD were annotated with ChaSen†23, while YACIS with MeCab†24. However, ChaSen and MeCab in their default settings use the same *ipadic* dictionary†25. Although there are some differences in the way each system disambiguates parts of speech, the same dictionary base makes it a good comparison of ipadic annotations on three different corpora (small JENAAD, larger jBlogs and large YACIS).

The statistics of POS distribution is similar between the pairs YACIS–JENAAD ($\rho = 1.0$ in Spearman's rank setting correlation test) and YACIS–jBlogs ($\rho = 0.96$).

5.2 Comparison of Emotion Corpora

In this section we compared sentiment-related annotations on emotion corpora. In particular we made three comparisons. Firstly, we compared positive and negative annotations on YACIS [32] and KNB [45]. Secondly, we compared the corpus by Minato et al. [51] with YACIS and Emotive Expression Database [39].

Firstly, we compared YACIS and KNB. The KNB corpus was annotated mostly for the needs of sentiment analysis and therefore does not contain any information on specific emotion classes. However, it is annotated with emotion valence for different categories valence can be expressed in Japanese, such as *emotional attitude* (e.g., "to feel sad about X"=negative, "to like X"=positive), *opinion* (e.g., "X is wonderful"=positive, "not

†22 We do not get into a detailed discussion on differences in performance between POS taggers for different languages, neither the discussion on whether the same POS labels (noun, verb, adjective, etc.) represent similar concepts among different languages (see for example [56] or [57] for this discussion). These two discussions, although important, are beyond the scope of this paper.

†23 http://chasen.naist.jp/

†24 http://sourceforge.jp/projects/mecab/

†25 http://sourceforge.jp/projects/ipadic/

Table 5 Comparison of numbers and ratios of positive and negative sentences between KNB and YACIS.

		positive	negative	ratio
KNB	emotional attitude	317	208	1.52
	opinion	489	289	1.69
	merit	449	264	1.70
	acceptation or rejection	125	41	3.05
	event	43	63	0.68
	sum	1,423	865	1.65
YACIS	only	22,381,992	12,837,728	1.74
	only+ mostly	23,753,762	13,605,514	1.75

convinced to X"=negative), or *positive/negative event* (e.g., "X broke down"=negative, "X was awarded"=positive). We compared the ratios of sentences expressing positive valence to the ones expressing negative valence. The comparison was made for all KNB valence categories separately and as a sum. In YACIS there is no additional sub-categorization of valence types, but we used in the comparison ratios of sentences with only positive/negative valence (expressions of emotions appearing in the sentence were of only positive valence) and including the mostly positive/negative sentences (situations when one sentence contains numerous expressions, majority of which is either positive or negative). The comparison is presented in Table 5. In KNB for all valence categories except one the ratio of positive to negative sentences was biased in favor of positive sentences. Moreover, for most cases, including the ratio taken from the sums of sentences, the ratio was similar to the one in YACIS (around 1.7 in favor of positive contents). Although the scale of compared corpora (number of compared sentences) differ greatly, the fact that the ratio remains similar across the two corpora of different sizes could suggest that the Japanese express in blogs more positive than negative emotions.

Next, we compared YACIS [32], the corpus created by Minato et al. [51] and the Emotive Expression Database (EED) [39]. Sentiment annotations in YACIS were based on the Emotive Expression Database (and thus indirectly on Nakamura's Emotive Expression Dictionary [54]) and share the same emotion labeling. On the other hand, emotion classes used in Minato et al. differ slightly to those used in YACIS and Nakamura's dictionary. For example, Minato et al. use class name "hate" to describe what in YACIS is called "dislike". Moreover, they have no classes such as "excitement", "relief" or "shame". On the other hand, They use class name "respect", which is not present in YACIS. To make the comparison possible and fair we excluded all emotion classes not appearing in all three corpora and unified all class names. The results are summarized in Table 6. A medium correlation was observed between YACIS and Minato et al. (ρ=0.57). Finally, a strong correlation was observed between Minato et al. and Nakamura (ρ=0.93), which is the most interesting observation. Both Minato et al. and Nakamura are in fact dictionaries of emotive expressions. The fact that they strongly correlate suggests that for the compared emotion classes there could be a tendency in the language to create more expressions to describe some emotions rather than the other (dislike, joy and fondness are often some of the most frequent emotion classes). Further and thorough analysis of

Table 6 Comparison of number of emotive expressions appearing in different corpora: Minato et al. [51], YACIS and EED/Nakamura's dictionary [54], with the results of Spearman's rank correlation test. Emotion classes marked with "*" were omitted to avoid unfair comparison.

	Minato et al.	YACIS	Nakamura
dislike	473	14,184,697	532
joy	339	22,100,500	224
fondness	274	13,817,116	197
sorrow	295	2,881,166	232
* relief	0	3,104,774	106
* excitement	0	2,833,388	269
anger	217	1,564,059	199
fear	209	4,496,250	147
* respect	119	0	0
surprise	31	3,108,017	129
* shame	0	952,188	65
	Minato et al. and Nakamura	Minato et al. and YACIS	YACIS and Nakamura
Spearman's ρ	0.93	0.57	0.25

this phenomenon in the future could reveal whether some languages have their specific emotional tendencies (for example, whether Americans or Polish tend to complain more than Japanese, etc.), which could put an end to or perhaps support some of the stereotypes about peoples.

6 Discussion on Availability of Corpora

An important issue in corpus development and application is the availability of corpora. A corpus could be a state of the art, however, if it is not available for other researchers its value dramatically decreases. In this section we discuss the availability of the corpora mentioned in this paper.

In the analysis of availability of the corpora we looked at three ways a corpus could be available. Firstly, a corpus could be open to the public and available for free download (either under an Open Source or similar license or under no conditions). A corpus available this way can be easily obtained by third party researchers and quickly applied to other research. Thus we consider open access as the most useful way of making a corpus available. Secondly, a corpus could be available after obtaining an agreement with the corpus developers. The "agreement" refers here to the further three situations: 1) personally contacting the corpus developers and asking for the permission to use the corpus; 2) applying for the permission to use the corpus through a predetermined and fixed application process (for example, registering the names of the future corpus users in a database, registering IP addresses of the computers on which the corpus will be further processed, etc.); 3) retail (when the license to use the corpus is sold to future users)[†26]. Finally, an important element in the process of making a corpus available is providing an online interface to search through the corpus. This is important in research not demanding massive corpus querying, such as numerous linguistic or corpus linguistic studies, or Web mining methods optimized for processing time, such as the one by Turney and Littman [22].

The availability of all corpora described in section 4, analyzed according to the above guidelines is compared in Table 7.

In general, large scale corpora have higher tendency to be widely available. Especially the WaCky project has contributed greatly to the creation of large scale corpora in various languages and making them widely available. The Google Books corpus is also available to the public, which makes it the largest freely available corpus today. Also a small number of emotion corpora are widely available as well, both for English (MPQA) and for Japanese (KNB, EED).

Most of the corpora are available after obtaining agreement with the corpus developers. In many cases the corpus is provided by the creators free of charge. The creators of emotion corpora tend to provide their work free of charge, whereas in some cases of the large scale corpora, the users need to pay for the license to use the corpus. When it comes to the online interface, very few corpora are equipped with this option. Almost none of the corpora freely available for download has online interface. This is most probably due to the fact that any corpus analysis can be done locally after downloading it.

As the corpora which do have the online interface we counted also Google 1T and Microsoft N-gram corpus, since these corpora are in fact the data indexed by the search engines. Although the search engine indexes and the corpora do not have the same coverage (corpus version tends to be smaller), the search engine results can be considered as larger and frequently updated version of the corpus.

Almost none of the emotion corpora, has online interface. The only two exceptions are KNB and YACIS. The former is provided with a locally accessible easy to use interface allowing navigating through the corpus contents. On the other hand, YACIS has an online interface similar to a standard search engine (Google or Bing) based on Apache Solr engine. Unfortunately, the engine does not allow searching the annotated contents (morphological annotations, emotion annotations, etc.), and the search is based on the tokenized version of the

[†26] We included non payable and payable agreements in one category to avoid an impression that a free of charge corpus is better than a payable one. We believe that corpora should be compared by objective features related to research applicability and financial context should not influence the impression of the corpus.

Table 7 Comparison of different corpora, ordered arbitrary by availability and size.

corpus	scale	language	type	availability		
name	(in words)		and	free	retail or user	online
			domain	download	agreement	interface
Google Books	155 billion	English	large scale/books/n-gram	○	○	○
National Corpus of Polish [30]	1 billion	Polish	large scale	○	○	○
ukWaC [28]	2 billion	English	large scale/Web	○	○	×
PukWaC [28]	2 billion	English	large scale/Web	○	○	×
itWaC [7] [28]	2 billion	Italian	large scale/Web	○	○	×
deWaC [28]	1.7 billion	German	large scale/Web	○	○	×
frWaC [28]	1.6 billion	French	large scale/Web	○	○	×
EED [54]	2 K	Japanese	emotion	○	○	×
KNB [45]	67 K	Japanese	blog/emotion	○	×	○
MPQA [43]	20 K	English	news/emotion	○	×	×
JpWaC [6]	400 million	Japanese	medium scale/Web	○	○	×
Google 1T	1 trillion	English	large scale/Web/n-gram	×	○	○
Microsoft N-gram	1.4 trillion	English	large scale/Web/n-gram	×	○	○
YACIS [11]	5.6 billion	Japanese	large scale/blog/emotion	×	○	○
Corpus Brasiliero [9]	1 billion	Brazilian Portuguese	large scale	×	○	○
BiWeC [15]	5.5 billion	English	large scale/Web	×	○	×
Gigaword [29]	2 billion	Hungarian	large scale	×	○	×
Ren-CECps1.0 [42]	878 K	Chinese	blog/emotion	×	○	×
Aman&Szpak. [40]	75 K	English	blog/emotion	×	○	×
Das&Bandyo. [46]	12,149 sentences	Bengali	blog/emotion	×	○	×
Liu&Curran [17]	10 billion	English	large scale/Web	×	×	×
jBlogs [6]	62 million	Japanese	medium scale/Web	×	×	×
WKEC [47]	400 K	Japanese	blog/emotion	×	×	×
Mishne [50]	815 K blog posts	English	blog/emotion	×	×	×
Minato [52]	14 K	Japanese	emotion	×	×	×

corpus.

7 Conclusions

In this paper we presented a survey on natural language corpora. Natural language corpora are crucial for many Software Engineering applications, such as part-of-speech taggers, dependency parsers, dialog systems or sentiment analysis software. Many of such applications are based on small scale resources. Recently there have been several initiatives to create large-scale Web-based corpora which could enhance such applications. We compared several such natural language corpora created for different languages, analyzed their distinctive features and the amount of additional annotations. In the survey on Web-based corpora we firstly focused on large-scale corpora containing billion words or more. Secondly, we focused on emotion corpora, widely applied in training sentiment analysis and opinion mining systems.

Next, we performed a detailed comparison of those corpora which are widely available for analysis or their descriptions contain enough details for thorough comparison. We analyzed three corpora of different languages (Japanese, British English and Italian), but comparable size (YACIS, ukWaC and itWaC, all containing 1 bil. words or more), and three other corpora of the same language (Japanese), but different sizes (large YACIS, medium jBlogs and small JENAAD). The comparison revealed interesting observations. The three corpora in Japanese, although different in size, showed similar POS distribution, whereas for other languages, although the corpora were comparable in size, the POS distribution differed greatly. We plan to investigate these differences in more detail in the future.

As for emotion corpora, the comparison between four corpora (YACIS, KNB, the corpus by Minato et al. [51], and Emotive Expression Database by Ptaszynski et al. [39]) showed similarities in the ratio of expressions of positive to negative emotions on small and large corpora. We also observed a high correlation between two different emotive expression dictionaries.

Natural language corpora have many practical Software Engineering applications. For example, Ptaszynski et al. [37] used YACIS to extract a random sample from the corpus to evaluate a system for affect analysis of emoticons. Later, Ptaszynski et al. [35] used YACIS to automatically construct a robust ontology of emotion objects. Finally, Rzepka and colleagues [34][33] recently used YACIS in their task for emotional and moral consequence retrieval. As for other examples natural language corpora (especially large-scale corpora) can be helpful with, Liu and Curran [17] used their 10-billion-word corpus for tasks such as spelling correction and thesaurus extraction. Turney and Littman [22] showed that large scale corpora can 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).

Acknowledgment

This research was partially supported by (JSPS) KAKENHI Grant-in-Aid for Scientific Research (Project number: 24600001) and (JSPS) KAKENHI Grant-in-Aid for Scientific Research (Project number: 24650061).

The authors wish to thank the following researchers for their help and advices during during preparing this survey (alphabetical order): Dipankar Das, Tomaž Erjavec, Pavel Rychly, Tony Berber Sardinha, Stan Szpakowicz.

References

[1] Charniak, E., Blaheta, D., Ge, N., Hall, K., Hale, J. and Johnson, M. : BLLIP 1987-89 WSJ Corpus Release 1, *Linguistic Data Consortium*, Philadelphia, 2000.

[2] Mainichi Shinbun CD, http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html (Retrieved: 2013.10.21)

[3] Ulrich, J., Murray, G. and Carenini, G. : A Publicly Available Annotated Corpus for Supervised Email Summarization, in *AAAI08 EMAIL Workshop*, Chicago, USA, 2008.

[4] Corpus of Spontaneous Japanese, http://www.ninjal.ac.jp/english/products/csj/ (Retrieved: 2013.10.21)

[5] Aozora Bunko, http://www.aozora.gr.jp/ (Retrieved: 2013.10.21))

[6] Erjavec, S., Erjavec, T. and Kilgarriff, A. : A web corpus and word sketches for Japanese, *Journal of Natural Language Processing* Vol. 15, No. 2 (2008), pp. 137–159.

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

[8] Kudo, T. and Kazawa, H. : Japanese Web N-gram Version 1, http://www.ldc.upenn.edu/Catalog/CatalogEntry.js?catalogId=LDC2009T08 (Retrieved: 2013.10.21)

[9] Sardinha, T. B., Moreira Filho, J. L. and Alambert, E. : The Brazilian Corpus, American Association for Corpus Linguistics, Edmonton, Canada,2009.

[10] Sasaki, Y., Isozaki, H., Taira, H., Hirao, T., Kazawa, H., Suzuki, J., Kokuryo, K. and Maeda, E. : SAIQA: A Japanese QA System Based on a Large-Scale Corpus [in Japanese], *IPSJ SIG Notes*, Vol. 2001, No. 86 (2001), pp. 77–82.

[11] Ptaszynski, M., Dybala, P., Rzepka, R., Araki, K. and Momouchi, Y. : YACIS: A Five-Billion-Word Corpus of Japanese Blogs Fully Annotated with Syntactic and Affective Information, in *Proceedings of The AISB/IACAP World Congress 2012 in Honour of Alan Turing, 2nd Symposium on Linguistic and Cognitive Approaches To Dialog Agents* (*LaCATODA 2012*), 2012, pp. 40–49.

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

[13] Zipf, G. K. : *The Psychobiology of Language*, Houghton-Mifflin, 1935.

[14] Zipf, G. K. : *Human Behavior and the Principle of Least Effort*, Addison-Wesley, 1949.

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

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

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

[18] Billsus, D. and Pazzani, M. : A hybrid user model for news story classification, in *Proceedings of the Seventh International Conference on User Modeling* (*UM'99*), 1999.

[19] Nahm, U. Y. and Mooney, R. J. : A mutually beneficial integration of data mining and information extraction, in *Proceedings of the Seventeenth National Conference on Artificial Intelligence* (*AAAI-00*), 2000.

[20] Ptaszynski, M., Dybala, P., Shi, W., Rzepka, R. and Araki, K. : A System for Affect Analysis of Utterances in Japanese Supported with Web Mining, *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, Special Issue on Kansei Retrieval, Vol. 21, No. 2 (2009), pp. 30–49 (194–213).

[21] Nitta, T., Masui, F., Ptaszynski, M., Kimura, Y., Rzepka, R. and Araki, K. : Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization, in *Proceedings of the 6th International Joint Conference on Natural Language Processing* (*IJCNLP 2013*), 2013.

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

[23] Kilgarriff, A. : Googleology is Bad Science, Last Words, *Computational Linguistics*, Vol. 33, No. 1 (2007), pp. 147–151.

[24] Brants, T. and Franz, A. : Web 1T 5-gram Version1, Linguistic Data Consortium, Philadelphia, 2006, http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html (Retrieved: 2013.10.22).

[25] Wang, K., Thrasher, C., Viegas, E., Li, X. and Hsu, B. : An Overview of Microsoft Web N-gram Corpus and Applications, in *Proceedings of the NAACL HLT 2010*, pp. 45–48.

[26] Yu, L. C., Wu, C. H., Philpot, A. and Hovy, E. H. : OntoNotes: sense pool verification using Google N-gram and statistical tests, in *Proceedings of OntoLex Workshop*, 2007.

[27] Andrew, C. and Fette, I. : Memory-based context-sensitive spelling correction at web scale, in *Proceedings of IEEE Sixth International Conference on Machine Learning and Applications* (*ICMLA*), 2007.

[28] Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. : The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora, *Language Resources and Evaluation*, Vol. 43, No. 3 (2009), pp. 209–226.

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

[30] Głowińska, K. and Przepiórkowski, A. : The Design of Syntactic Annotation Levels in the National Corpus of Polish, in *Proceedings of LREC*, 2010.

[31] Maciejewski, J., Ptaszynski, M. and Dybala, P. : 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*), 2010, pp. 192–195.

[32] Ptaszynski, M., Rzepka, R., Araki, K. and Momouchi, Y. : Automatically Annotating A Five-Billion-Word Corpus of Japanese Blogs for Affect and Sentiment Analysis, in *Proceedings of the 3rd ACL Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (*WASSA 2012*), 2012, pp. 89–98.

[33] Komuda, R., Ptaszynski, M., Momouchi, Y., Rzepka, R. and Araki, K. : Machine Moral Development: Moral Reasoning Agent Based on Wisdom of Web-Crowd and Emotions, *Int. Journal of Computational Linguistics Research*, Vol. 1, Issue 3 (2010), pp. 155–163.

[34] Rzepka, R. and Araki, K. : What Statistics Could Do for Ethics? - The Idea of Common Sense Processing Based Safety Valve, AAAI Fall Symposium on Machine Ethics, Technical Report FS-05-06, 2005, pp. 85–87.

[35] Ptaszynski, M., Rzepka, R., Araki, K. and Momouchi, Y. : A Robust Ontology of Emotion Objects, in *Proceedings of The Eighteenth Annual Meeting of The Association for Natural Language Processing* (*NLP-2012*), 2012, pp. 719–722.

[36] Kilgarriff, A., Rychly, P., Smrž, P. and Tugwell, D. : The Sketch Engine, in *Proc. EURALEX*, 2004, pp. 105–116.

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

[38] Matsumoto, K., Konishi, Y., Sayama, H. and Ren, F. : Analysis of Wakamono Kotoba Emotion Corpus and Its Application in Emotion Estimation, *International Journal of Advanced Intelligence*, Vol. 3, No. 1 (2001), pp. 1–24.

[39] Ptaszynski, M., Dybala, P., Rzepka, R. and Araki, K. : Affecting Corpora: Experiments with Automatic Affect Annotation System - A Case Study of the 2channel Forum -, in *Proceedings of the Conference of the Pacific Association for Computational Linguistics* (*PACLING-09*), 2009, pp. 223–228.

[40] Aman, S. and Szpakowicz, S. : Identifying Expressions of Emotion in Text, in *Proceedings of the 10th International Conference on Text, Speech, and Dialogue* (*TSD-2007*),V. Matousek, P. Mautner (eds.), Plzeň, Czech Republic, Lecture Notes in Computer Science (LNCS) 4629, Springer, 2007, pp. 196–205.

[41] Aman, S. and Szpakowicz, S. : Using Roget's

Thesaurus for Fine-grained Emotion Recognition, in *Proceedings of the Third International Joint Conference on Natural Language Processing* (*IJCNLP 2008*), Hyderabad, India, 2008, pp. 296–302.

[42] Quan, C. and Ren, F. : A blog emotion corpus for emotional expression analysis in Chinese, *Computer Speech & Language*, Vol. 24, Issue 4 (2010), pp. 726–749.

[43] Wiebe, J., Wilson, T. and Cardie, C. : Annotating expressions of opinions and emotions in language, *Language Resources and Evaluation*, Vol. 39, Issue 2-3 (2005), pp. 165–210.

[44] Wilson, T. and Wiebe, J. : Annotating Attributions and Private States, in *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II*, 2005, pp. 53–60.

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

[46] Das, D. and Bandyopadhyay, S. : Labeling Emotion in Bengali Blog Corpus - A Fine Grained Tagging at Sentence Level, in *Proceedings of the 8th Workshop on Asian Language Resources*, 2010, pp. 47–55.

[47] Matsumoto, K., Kita, K. and Ren, F. : Emotion Estimation from Sentence Using Relation between Japanese Slangs and Emotion Expressions, in *The 26th Pacific Asia Conference on Language, Information, and Computation* (*PACLIC2012*), 2012, pp. 343–350.

[48] Ekman, P. : An Argument for Basic Emotions, *Cognition and Emotion*, Vol. 6 (1992), pp. 169–200.

[49] Fischer, K. W., Shaver, P. R. and Carnochan, P. : A Skill Approach to Emotional Development: From Basic-to Subordinate-category Emotions,: Damon,W.(Ed),*Child development today and tomorrow, The Jossey-Bass social and behavioral science series*, San Francisco, CA, 1989, pp. 107–136.

[50] Mishne, G. : Experiments with Mood Classification in Blog Posts, in *Style2005: the 1st Workshop on Stylistic Analysis of Text for Information Access, SIGIR 2005.*

[51] Minato, J., Bracewell, D. B., Ren, F. and Kuroiwa, S. : *Statistical Analysis of a Japanese Emotion Corpus for Natural Language Processing*, LNCS Vol. 4 114, 2006.

[52] Minato, J., Matsumoto, K., Ren, F., Tsuchiya, S. and Kuroiwa, S. : Evaluation of Emotion Estimation Methods Based on Statistic Features of Emotion Tagged Corpus, *International Journal of Innovative Computing*, Information and Control, Vol. 4, No. 8 (2008), pp. 1931–1941.

[53] Hiejima, I. : *A short dictionary of feelings and emotions in English and Japanese*, Tokyodo Shuppan, 1995.

[54] Nakamura, A. : *Kanjō hyōgen jiten* [Dictionary of Emotive Expressions] (in Japanese), Tokyodo Publishing, Tokyo, 1993.

[55] Utiyama, M. and Isahara, H. : Reliable Measures for Aligning Japanese-English News Articles and Sentences, in *Proceedings of ACL-2003*, pp. 72–79.

[56] Hopper, P. and Thompson, S. : The Iconicity of the Universal Categories 'Noun' and 'Verbs', *Typological Studies in Language: Iconicity and Syntax*, Haiman, J. (ed.), Vol. 6, Amsterdam: John Benjamins Publishing Company, 1985, pp. 151–83.

[57] Broschart, J. : Why Tongan does it differently: Categorial Distinctions in a Language without Nouns and Verbs, *Linguistic Typology*, Vol. 1, No. 2 (1997), pp. 123–165.

[58] Backhouse, A. E. : Have all the adjectives gone?, *Lingua*, Vol. 62, Issue 3 (1984), pp. 169–186.

Michal Ptaszynski

Michal Ptaszynski was born in Wroclaw, Poland in 1981. He received the MA degree from the University of Adam Mickiewicz, Poznan, Poland, in 2006, and PhD in Information Science and Technology from Hokkaido University, Japan in 2011. In years 2001-2013 he was a JSPS Post-doctoral Research Fellow at the High-Tech research Center, Hokkai-Gakuen University, Japan. Currently he is an assistant professor at the Kitami Institute of Technology. His research interests include natural language processing, dialogue processing, affect analysis, sentiment analysis, HCI, and information retrieval. He is a member of the ACL, the AAAI, the IEEE, the HUMAINE, the AAR, the SOFT, the JSAI, and the ANLP.

Rafal Rzepka

Rafal Rzepka received the MA degree from the University of Adam Mickiewicz, Poznan, Poland, in 1999, and the PhD degree from Hokkaido University, Japan, in 2004. Currently, he is an assistant professor in the Graduate School of Information Science and Technology at Hokkaido University. His research interests include natural language processing, Web mining, common sense retrieval, dialogue processing, language acquisition, affect analysis, and sentiment analysis. He is a member of the AAAI, the ACL, the JSAI, the IPSJ, the IEICE, the JCSS, and the ANLP.

Satoshi Oyama

Satoshi Oyama is an associate professor in the Graduate School of Information Science and Technology, Hokkaido University, Japan. He received his B.Eng., M.Eng., and Ph.D. degrees from Kyoto University in 1994, 1996, and 2002, respectively. He was a research fellow of the Japan Society for the Promotion of Science from 2001 to 2002. He was an assistant professor in the Graduate School of Informatics at Kyoto University from 2002 to 2009. He was a visiting assistant professor in the Department of Computer Science at Stanford University from 2003 to 2004. His research interests include machine learning, data mining, information retrieval, crowdsourcing, and human computation. He is a member of IEEE, ACM, AAAI, IEICE, IPSJ, JSAI, and DBSJ.

Masahito Kurihara

Masahito Kurihara is a professor in the Graduate School of Information Science and Technology, Hokkaido University, Japan, where from 2010 to 2013 academic years, he worked as dean. He received his B.Eng., M.Eng., and Ph.D. degrees from Hokkaido University in 1978, 1980, and 1986, respectively. His research interests include theoretical computer science, software science, and automated reasoning in artificial intelligence. He is a member of IPSJ, IEICE, JSAI, and SOFT.

Kenji Araki

Kenji Araki received the BE, ME, and PhD degrees in electronics engineering from Hokkaido University, Sapporo, Japan, in 1982, 1985, and 1988, respectively. In April 1988, he joined Hokkai-Gakuen University, Sapporo, Japan, where he was a professor. He joined Hokkaido University in 1998 as an associate professor in the Division of Electronics and Information Engineering and became a professor in 2002. Presently, he is a professor in the Division of Media and Network Technologies at Hokkaido University. His research interests include natural language processing, spoken dialogue processing, machine translation, and language acquisition. He is a member of the AAAI, the IEEE, the JSAI, the IPSJ, the IEICE, and the JCSS.