YACIS: A Five-Billion-Word Corpus of Japanese Blogs Fully Annotated with Syntactic and Affective Information

Michal Ptaszynski[1] **Pawel Dybala**[2] **Rafal Rzepka**[3] **Kenji Araki**[4] and **Yoshio Momouchi**[5]

Abstract. This paper presents YACIS, a new fully annotated large scale corpus of Japanese language. The corpus is based on blog entries from Ameba blog service. The original structure (blog post and comments) is preserved, thanks to which semantic relations between posts and comments are maintained. The corpus is annotated with syntactic (POS, dependency parsing, etc.) and affective (emotive expressions, emoticons, valence, etc.) information. The annotations are evaluated in a survey on over forty respondents. The corpus is also compared to other existing corpora, both large scale and emotion related.

1 INTRODUCTION

Text corpora are some of the most vital linguistic resources in natural language processing (NLP). These include newspaper corpora [1], conversation corpora or corpora of literature[6]. Unfortunately, comparing to major world languages, like English, there are few large corpora available for the Japanese language. Moreover, grand majority of them is based on newspapers, or legal documents[7]. These are usually unsuitable for the research on sentiment analysis and emotion processing, as emotions and attitudes are rarely expressed in this kind of texts. Although there exist conversation corpora with speech recordings, which could become useful in such research[8], due to the difficulties with compilation of such corpora they are relatively small. Recently blogs have been recognized as a rich source of text available for public. Blogs are open diaries in which people encapsulate their own experiences, opinions and feelings to be read and commented by other people. Because of their richness in subjective and evaluative information blogs have come into the focus in sentiment and affect analysis [2, 3, 4, 5]. Therefore creating a large blog-based emotion corpus could become a solution to overcome both problems, of the lack in quantity of corpora and their applicability in the research on sentiment analysis and emotion processing. However, there have been only a few small (several thousand sentences) Japanese emotion corpora developed so far [2]. Although there exist medium scale Web-based corpora (containing several million words), such as JpWaC [6] or jBlogs [7], access to them is usually allowed only from the Web interface, which makes additional annotations (parts-of-speech, dependency structure, deeper affective information, etc.) difficult. Furthermore, although there exist large resources, like Google N-gram Corpus [8], the textual data sets in such resources are short (up to 7-grams) and do not contain any contextual information. This makes them unsuitable for emotion processing research, since most of contextual information, so important in expressing emotions [9], is lost. Therefore we decided to create a new corpus from scratch. The corpus was compiled using procedures similar to the ones developed in the WaCky initiative [10], but optimized to mining only one blog service (Ameba blog, http://ameblo.jp/, later referred to as Ameblo). The compiled corpus was fully annotated with syntactic (POS, lemmatization, dependency parsing, etc.) and affective information (emotive expressions, emotion classes, valence, etc.).

The outline of the paper is as follows. Section 2 describes the related research in large scale corpora and blog emotion corpora. Section 3 presents the procedures used in compilation of the corpus. Section 4 describes tools used in corpus annotation. Section 5 presents detailed statistical data and evaluation of the annotations. Finally the paper is concluded and applications of the corpus are discussed.

2 RELATED RESEARCH

In this section we present some of the most relevant research related to ours. There has been no billion-word-scale corpus annotated with affective information before. Therefore we needed to divide the description of the related research into “Large Scale Corpora” and “Emotion Corpora”.

2.1 Large-Scale Web-Based Corpora

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. [11] 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, the corpus 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%). 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 rather 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

[1] Hokkai-Gakuen University, Japan, email: ptaszynski@hgu.jp
[2] Otaru University of Commerce, Japan, email: paweldybala@res.otaru-uc.ac.jp
[3] Hokkaido University, Japan, email: kabura@media.eng.hokudai.ac.jp
[4] Hokkaido University, Japan, email: araki@media.eng.hokudai.ac.jp
[5] Hokkai-Gakuen University, Japan, email: momouchi@eli.hokkai-s-u.ac.jp
[6] http://www.aozora.gr.jp/
[7] http://www-nagao.kuee.kyoto-u.ac.jp/NLP_Portal/lr-cat-e.html
[8] http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/public/

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], 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 [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 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 [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 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 [17]).

The latter, followed by Baroni and Ueyama [7], indicate at least two types of research dealing with large-scale corpora. One is using popular search engines, such as Google[9] or Yahoo![10]. In such research 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 [18]. 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[11]. However, this kind of research is inevitably constrained with limitations of the search engine's API. Pomikálek et al. [15] 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 [19] 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 World Wide Web, and argues that it is the optimal way of utilizing the Internet contents for research in linguistics and computational linguistics.

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 [15, 19]. However, Google has presented two large corpora. One is the "Web 1T (trillion) 5 gram" corpus [47] published in 2006. It is estimated to contain one trillion of tokens extracted 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"[12] published recently 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[13], unfortunately only in the n-gram form (no context larger than 5-gram).

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. 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) [7, 10] 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 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 **Bi**g **W**eb **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[14] 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 [24], Brazilian Portuguese [46] or Polish [23].

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.

Srdanović Erjavec et al. [20] notice this lack of freely available large corpora for Japanese. They used WAC (Web As Corpus) Toolkit[15], developed under the WaCky initiative, and Kilgariff et al.'s [21] Sketch Engine, a tool for thesauri generation from large scale corpora (applied also for English in [15]). They gathered **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[16]. However, they faced several problems. Firstly, they had to normalize the character

[9] http://www.google.com
[10] http://www.yahoo.com
[11] In 2004 AltaVista (http://www.altavista.com/) has become a part of Yahoo!.
[12] http://googlebooks.byu.edu/
[13] http://books.google.com/ngrams/datasets
[14] http://wacky.sslmit.unibo.it/
[15] http://www.drni.de/wac-tk/
[16] languages like Chinese, Japanese or Korean are encoded using 2-bite characters.

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

corpus name	scale (in words)	language	domain	annotation
Liu&Curran [17]	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 [15]	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) [10]	2 billion	English	whole Web (.uk domain)	POS, lemma, dependency parsing;
itWaC [7, 10]	2 billion	Italian	whole Web (.it domain)	POS, lemma;
Gigaword [24]	2 billion	Hungarian	whole Web (.hu domain)	tokenization, sentence segmentation;
deWaC [10]	1.7 billion	German	whole Web (.de domain)	POS, lemma;
frWaC [10]	1.6 billion	French	whole Web (.fr domain)	POS, lemma;
Corpus Brasiliero [46]	1 billion	Brazilian Portuguese	multi-domain (newspapers, Web, talk transcriptions)	POS, lemma;
National Corpus of Polish [23]	1 billion	Polish	multi-domain (newspapers, literature, Web, etc.)	POS, lemma, dependency parsing, named entities, word senses;
JpWaC [20]	400 million	Japanese	whole Web (.jp domain)	tokenization, POS, lemma;
jBlogs [20]	62 million	Japanese	Blogs (Ameba, Goo, Livedoor, Yahoo!)	tokenization, POS, lemma;

Table 2. Detailed 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 [20]	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 [2]	66,952	249	4,186	450 kB	Blogs (written by students exclusively for the purpose of the research);
Minato et al. [29]	14,195	1	1,191	[not revealed]	Dictionary examples (written by dictionary authors);

encoding for all web pages[17] (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 [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, namely *kaomoji*[18]. 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 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. [34].

Apart from the above Kawahara and Kurohashi [27] claim the creation of a large, about two-billion-word corpus. However, detailed description of this corpus is not available.

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

[18] For more detailed description of Japanese emoticons, see [34].

Finally, Okuno Yoo and Sasano Manabu 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 [45]. No information on the corpus is yet available except methods of development, tools (tokenization by MeCab, a POS tagger for Japanese) and its size (1TB).

2.2 Emotion and Blog Corpora

The existing emotion corpora are mostly of limited scale and are annotated manually. Below we compare some of them. As an interesting remark, five out of six were extracted from blogs.

Quan and Ren in 2010 [5] created a Chinese emotion blog corpus **Ren-CECps1.0**. They collected 500 blog articles from various Chinese blog services, such as sina blog (http://blog.sina.com.cn/) or qq blog (http://blog.qq.com/). The articles were annotated with a variety of information, such as emotion class, emotive expressions or valence. Although the syntactic annotations were simplified to tokenization and POS tagging, the corpus is comparable to YACIS in the overall variety of annotations. The motivation for Quan and Ren is also similar - the lack of large scale corpora for sentiment analysis and emotion processing research in Chinese (in our case - Japanese). Wiebe and colleagues [38, 39] created the **MPQA** corpus of news articles. It contains 10,657 sentences in 535 documents. The annotations include emotive expressions, valence, intensity, etc. However, Wiebe et al. focused mostly on sentiment and subjectivity analysis, and they did not include annotations of emotion classes. Hashimoto et al. [2] developed the **KNB** corpus of Japanese blogs. The corpus contains about 67 thousand words in 249 blog articles. Despite its small scale, the corpus proposes a good standard for preparation

Table 3. Comparison of emotion corpora ordered by the amount of annotations.

corpus name	scale (in sentences / docs)	language	annotated affective information					syntactic annotations
			emotion classes (standard)	emotive expressions	emotive/ non-emot.	valence/ activation	emotion intensity	
YACIS	354 mil. / 13 mil.	Japanese	10 (language and culture based)	○	○	○/○	○	T,POS,L,DP,NER;
Ren-CECps1.0	12,724 / 500	Chinese	8 (Yahoo! news)	○	○	○/×	○	T,POS;
MPQA	10,657 / 535	English	none (no standard)	○	○	○/×	○	T,POS;
KNB	4,186 / 249	Japanese	none (no standard)	○	×	○/×	×	T,POS,L,DP,NER;
Minato et al.	1,191 / 1	Japanese	8 (chosen subjectively)	○	○	×/×	×	POS;
Aman&Szpak.	5205 / 173	English	6 (face recognition)	○	○	×/×	○	×

of blog corpora for sentiment and affect-related studies. It contains all relevant syntactic annotations (POS, dependency parsing, Named Entity Recognition, etc.). It also contains sentiment-related information. Words and phrases expressing emotional attitude were annotated by laypeople as either positive or negative. A disadvantage of the corpus, except its small scale, is the way it was created. Eighty students were employed to write blogs for the need of the 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 in private. In YACIS we included all types of information contained in KNB, and also added more, especially in the affect-related annotations. Aman and Szpakowicz [4] created a small-scale English blog corpus in 2007. They focused not on syntactic, but on affect-related annotations. They were also some of the first to recognize the task of distinguishing between emotive and non-emotive sentences. ML-Ask, a system applied in annotation of YACIS was evaluated in this matter with high accuracy. Finally, Minato et al. [29] collected a 14,195 word / 1,191 sentence corpus. The corpus is a collection of dictionary examples from "A short dictionary of feelings and emotions in English and Japanese" [25]. It is a dictionary created for Japanese language learners. The sentence examples were mostly prepared by the dictionary author. Moreover, the dictionary does not propose any coherent emotion class list, but rather the emotion concepts are chosen subjectively. All the above corpora were annotated manually or semi-automatically. In our research we performed the first attempt to annotate affect on a large scale corpus automatically. We performed this with previously developed systems, thoroughly evaluated and based on a standardized emotion class typology.

3 YACIS CORPUS COMPILATION

The corpus (named **YACIS** Corpus, or **Y**et **A**nother **C**orpus of **I**nternet **S**entences) was assembled using data obtained automatically from the pages of Ameba Blog (www.ameblo.co.jp, below referred to as Ameblo). There were two main reasons for using Ameblo. Firstly, the users are mostly Japanese so the risk that the links may lead to pages written in a language other than Japanese is small. Secondly, Ameblo has a clear structure of HTML source code, which makes it easy to extract only posts and comments omitting the irrelevant contents, such as advertisements or menu links.

All the tools used for compiling this corpus were developed especially for the purpose of this research. Although there existed several other solutions, all of them were created for crawling the whole Web and included some parts irrelevant for crawling blog service urls like Ameblo (such as the detection of "robots.txt" file, which specifies that no robots should visit any URL from the domain, used for privacy protection), or parts that can be done more easily if the crawling domain is restricted to one blog service (such as HTML code boilerplate deletion). All these parts slow down the crawling process, and sometimes influence the corpus extraction (e.g., general rules for HTML code deletion are less precise than specific rules for deletion of the HTML code that appears in Ameblo). Therefore the available tools, very useful as they are, were insufficient for our needs. All our tools were written in C# and are operating under MS Windows systems.

We developed a simple but efficient web crawler designed to crawl exclusively Ameblo Web pages. The only pages taken into account were those containing Japanese posts (pages with legal disclaimers, as well as posts written in English and other languages were omitted). Initially we fed the crawler with 1000 links taken from Google (response to a query: 'site:ameblo.jp'). All the pages were saved to disk as raw HTML files (each page in a separate file) to be processed later. All of them were downloaded within three weeks between 3rd and 24th of December 2009. Next, we extracted all the posts and comments and divided them into sentences.

Although sentence segmentation may seem to be a trivial task it is not that easy when it comes to texts written by bloggers. People often use improper punctuation, e.g., the periods at the end of sentences are often omitted. In that case we assumed that if the given parts of text are separated by two
 tags (two markers of a new line) then those parts will be two separate sentences. This does not solve the problem in all cases. Therefore we rejoined previously separated parts if the first part ended with a coma or if the quotation marks or parenthesis were opened in the first part and closed in second.

Unfortunately, these modifications were still not perfect and in several cases parts of the text remained not split while others were segmented erroneously. One of the possible improvements was to take into consideration emoticons. We observed that if an emoticon is present in the sentence it usually appears at the end of it. Even in the cases the emoticon did not appear on the very end of the sentence, it still separated two clauses of a different meaning. Moreover, the meaning of the emoticon was always bound with the clause preceding it. This suggested separating sentences after emoticons. To do that we used CAO emoticon analysis system developed by Ptaszynski, et al. [34]. Observations showed this coped with most of the remaining sentence segmentation errors. In a random 1000 sentence sample, less than 1% remained erroneously separated. Analysis of errors showed these were sentences separated by blog authors in a non-standard way and without any particular rule. However, since such cases did not exceed a 5% border of statistical error we considered them an agreeable error.

Currently the data is stored in modified-XML format. Although it looks like XML it does not comply with all XML standards due to the presence of some characters forbidden by XML specification, such as apostrophes (') or quotation marks ("). Those modifications were made to improve the communication with natural language processing tools used in further processing of the corpus, such as a text

parser, part-of-speech analyzer (e.g., MeCab [41]), affect analysis system (ML-Ask [33]) and others. Each page was transformed into an independent XML block between <doc></doc> tags. Opening tag of the <doc> block contains three parameters: URL, TIME and ID which specify the exact address from which the given page was downloaded, download time and unique page number, respectively. The <doc> block contains two other tag types: <post> and <comments>. The <post> block contains all the sentences from the given post where each sentence is included between <s></s> tags. The block <comments> contains all comments written under given post placed between <cmt></cmt> tags which are further split into single sentences placed between <s></s> tags (as described above). An example XML structure of the corpus is represented in figure 1.

The corpus is stored in 129 text files containing 100 000 <doc> units each. The corpus was encoded using UTF-8 encoding. The size of each file varies and is between 200 and 320 megabytes. The size of raw corpus (pure text corpus without any additional tags) is 27.1 gigabytes. Other primary statistics of the corpus are represented in the table 4 below.

Table 4. General Statistics of YACIS Corpus

# of web pages	12,938,606
# of unique bloggers	60,658
average # of pages/blogger	213.3
# of pages with comments	6,421,577
# of comments	50,560,024
average # of comment/page	7.873
# of characters (without spaces)	28,582,653,165
# of characters (with spaces)	34,202,720,910
# of words	5,600,597,095
# of all sentences	354,288,529
# of sentences < 500 characters	353,999,525
# of sentences after correction of sentence segmentation errors	371,734,976
# of words per sentence (average)	15
# of characters per sentence (average)	77

As mentioned in Table 4, average sentence length is 28.17 Japanese characters. Kubota et al. [44] divide sentences in Japanese according to their intelligibility into: easily intelligible short sentences (up to 100 characters) and difficult long sentences (over 100 characters long). The sentences in our corpus fit in the definition of short sentences which means they are easily understandable. After exclusion of very long sentences (consisting of over 500 characters) the number of sentences does not change significantly and is 354,169,311 (99,96%) with an average length of 27.9 characters. This means the corpus is balanced in the length of sentences.

4 YACIS CORPUS ANNOTATION TOOLS

The corpus, in the form described in section 3 was further annotated with several kinds if information, such as parts-of-speech, dependency structure or affective information. The tools we used in particular are described in detail below.

4.1 Syntactic Information Annotation Tools

MeCab [41] is a standard morphological analyzer and parts-of-speech (POS) tagger for Japanese. It is trained using a large corpus on a Conditional Random Fields (CRF) discriminative model and uses a bigram Markov model for analysis. Prior to MeCab there were several POS taggers for Japanese, such as Juman[19] or ChaSen[20]. ChaSen and MeCab have many similarities in their structures. Both share the same corpus base for training and use the same default dictionary (ipadic[21] based on a modified IPA Part of Speech Tagset developed by the Information-Technology Promotion Agency of Japan (IPA)). However, ChaSen was trained on a Hidden Markov Model (generative model), a full probabilistic model in which first all variables are generated, and thus is slower than MeCab, based on a discriminative model, which focuses only on the target variables conditional on the observed variables. Juman on the other hand was developed separately from MeCab on different resources. It uses a set of handcrafted rules and a dictionary (jumandic) created on the basis of Kyoto Corpus developed by a Kurohashi&Kawahara Laboratory[22] at Kyoto University. Both MeCab and Juman are considerably fast, which is a very important feature when processing a large-scale corpus such as YACIS. However, there were several reasons to choose the former. MeCab is considered slightly faster when processing large data and uses less memory. It is also more accurate since it allows partial analysis (a way of flexible setting of word boundaries in non-spaced languages, like Japanese). Finally, MeCab is more flexible when using other dictionaries. Therefore to annotate YACIS we were able to use MeCab with the two different types of dictionaries mentioned above (ipadic and jumandic). This allowed us to develop POS tagging for YACIS with the two most favored standards in morphological analysis of Japanese today. An example of MeCab output is represented in figure 2 (the results were translated into English according to Francis Bond's "IPA POS code in Japanese and English"[23] developed as a standard for annotation of Japanese WordNet[24]).

Cabocha [42] is a Japanese dependency parser based on Support Vector Machines. It was developed by MeCab developers and is considered to be the most accurate statistical Japanese dependency parser. Its discriminative feature is using Cascaded Chunking Model, which makes the analysis efficient for the Japanese language. Other dependency parsers for Japanese, such as KNP[25] use statistical probabilistic models, which makes them inefficient for complex sentences with many clauses. Cascaded Chunking Model parses a sentence deterministically focusing on whether a sentence segment modifies a segment on its right hand side [42]. As an option, Cabocha uses IREX[26] (Information Retrieval and Extraction Exercise) standard for Named Entity Recognition (NER). We applied this option in the annotation process as well. An example of Cabocha output is represented in figure 2. Table 5 represents all tag types included in IREX.

Table 5. Named entity annotations included in the IREX standard.

<opening tag>...</closing tag>	**explanation**
<ORGANIZATION>... ...</ORGANIZATION>	organization or company name including abbreviations (e.g., Toyota, or Nissan);
<LOCATION>...</LOCATION>	mane of a place (city, country, etc.);
<PERSON>...</PERSON>	name, nickname, or status of a person (e.g., Lady Gaga, or "me", "grandson", etc.);
<ARTIFACT>...</ARTIFACT>	name of a well recognized product or object (e.g., Van Houtens Cocoa, etc.);
<PERCENT>...</PERCENT>	percentage or ratio (90%, 0.9);
<MONEY>...</MONEY>	currencies (1000 $, 100 ¥);
<DATE>...</DATE>	dates and its paraphrased extensions (e.g., "4th July", but also "next season", etc.)
<TIME>...</TIME>	hours, minutes, seconds, etc.

[19] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN
[20] http://chasen.naist.jp/hiki/ChaSen/
[21] http://sourceforge.jp/projects/ipadic/
[22] http://nlp.ist.i.kyoto-u.ac.jp/index.php
[23] http://sourceforge.jp/projects/ipadic/docs/postag.txt
[24] http://nlpwww.nict.go.jp/wn-ja/index.en.html
[25] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP
[26] http://nlp.cs.nyu.edu/irex/index-e.html

```
<doc url="http://ameblo.jp/capo-del-rosso/entry-000000.html" time="2009-12-05 21:11:46" id="2000001">
        <post>
                <s>今日から十月です。</s>
                [Its October from today.]
                <s>なんか、九月はいつもよりアッという間に過ぎたような気がするなぁ。</s>
                [I have a strange feeling September passed faster than usual.]
                ...
        </post>
        <comments>
                <cmt>
                        <s>色々と忙しいですね〜！</s>
                        [Oh, you've been busy, weren't you?]
                        ...
                </cmt>
                <cmt>
                        <s>お疲れヤマです(^O^)</s>
                        [Well done! Cheers for good work (^o^)]
                        ...
                </cmt>
        </comments>
</doc>
```

Figure 1. The example XML structure of the main blog corpus.

4.2 Affective Information Annotation Tools

Emotive Expression Dictionary [30] is a collection of over two thousand expressions describing emotional states collected manually from a wide range of literature. It is not a tool *per se*, but was converted into an emotive expression database by Ptaszynski et al. [33]. Since YACIS is a Japanese language corpus, for affect annotation we needed the most appropriate lexicon for the language. The dictionary, developed for over 20 years by Akira Nakamura, is a state-of-the art example of a hand-crafted emotive expression lexicon. It also proposes a classification of emotions that reflects the Japanese culture: 喜 *ki/yorokobi* (joy), 怒 *dō/ikari* (anger), 哀 *ai/aware* (sorrow, sadness, gloom), 怖 *fu/kowagari* (fear), 恥 *chi/haji* (shame, shyness), 好 *kō/suki* (fondness), 厭 *en/iya* (dislike), 昂 *kō/takaburi* (excitement), 安 *an/yasuragi* (relief), and 驚 *kyō/odoroki* (surprise). All expressions in the dictionary are annotated with one emotion class or more if applicable. The distribution of expressions across all emotion classes is represented in Table 6.

Table 6. Distribution of separate expressions across emotion classes in Nakamura's dictionary (overall 2100 ex.).

emotion class	nunber of expressions	emotion class	nunber of expressions
dislike	532	fondness	197
excitement	269	fear	147
sadness	232	surprise	129
joy	224	relief	106
anger	199	shame	65
		sum	**2100**

ML-Ask [31, 33] is a keyword-based language-dependent system for affect annotation on utterances in Japanese. It uses a two-step procedure: **1)** specifying whether an utterance is emotive, and **2)** annotating the particular emotion classes in utterances described as emotive. The emotive sentences are detected on the basis of *emotemes*, emotive features like: interjections, mimetic expressions, vulgar language, emoticons and emotive markers. The examples in Japanese are respectively: *sugee* (great!), *wakuwaku* (heart pounding), *-yagaru* (syntactic morpheme used in verb vulgarization), (^_^) (emoticon expressing joy) and '!', '??' (markers indicating emotive engagement). Emotion class annotation is based on Nakamura's dictionary. ML-Ask is also the only present system for Japanese recognized to implement the idea of Contextual Valence Shifters (CVS) [40] (words and phrases like "not", or "never", which change the valence of an evaluative word). The last distinguishable feature of ML-Ask is implementation of Russell's two dimensional affect model [36], in which emotions are represented in two dimensions: valence (positive/negative) and activation (activated/deactivated). An example of negative-activated emotion could be "anger"; a positive-deactivated emotion is, e.g., "relief". The mapping of Nakamura's emotion classes on Russell's two dimensions was proved reliable in several research [32, 33, 34]. With these settings ML-Ask detects

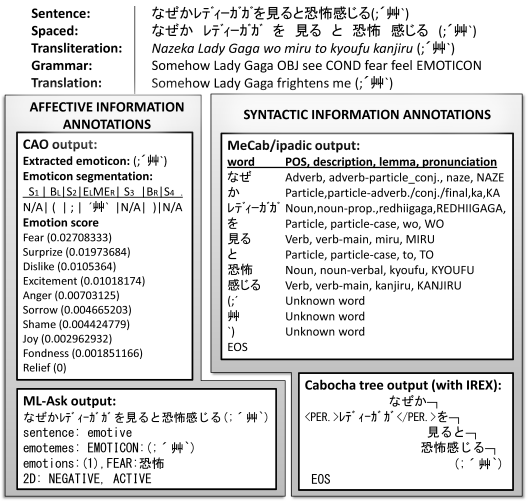

Figure 2. Output examples for all systems.

emotive sentences with a high accuracy (90%) and annotates affect on utterances with a sufficiently high Precision (85.7%), but low Recall (54.7%). Although low Recall is a disadvantage, we assumed that in a corpus as big as YACIS there should still be plenty of data.

CAO [34] is a system for affect analysis of Japanese emoticons, called *kaomoji*. Emoticons are sets of symbols used to convey emotions in text-based online communication, such as blogs. CAO extracts emoticons from input and determines specific emotions expressed by them. Firstly, it matches the input to a predetermined raw emoticon database (with over ten thousand emoticons). The emoticons, which could not be estimated with this database are divided into semantic areas (representations of "mouth" or "eyes"). The areas are automatically annotated according to their co-occurrence in the database. The performance of CAO was evaluated as close to ideal [34] (over 97%). In the YACIS annotation process CAO was used as a supporting procedure in ML-Ask to improve the overall performance and add detailed information about emoticons.

5 ANNOTATION RESULTS AND EVALUATION

5.1 Syntactic Information

In this section we present all relevant statistics concerning syntactic information annotated on YACIS corpus. Where it was possible we also compared YACIS to other corpora. All basic information concerning YACIS is represented in table 4. Information on the distribution of parts of speech is represented in table 7. We compared the two dictionaries used in the annotation (ipadic and jumandic) with other Japanese corpora (jBlogs, and JENAAD newspaper corpus) and in addition, partially to British and Italian Web corpus (ukWaC and itWaC, respectively). The results of analysis are explained below.

Ipadic vs Jumandic: There were major differences in numbers of each part-of-speech type annotations between the dictionaries. In most cases ipadic provided more specific annotations (nouns, verbs, particles, auxiliary verbs, exclamations) than jumandic. For example, in ipadic annotation there were nearly 2 billions of nouns, while in jumandic only about 1,5 billion (see table 7 and its graphical visualization in figure 3 for details). The reason for these differences is that both dictionaries are based on different approaches to part-of-speech disambiguation. Jumandic was created using a set of hand crafted syntactic rules and therefore in a corpus as large as YACIS there are situations where no rule applies. On the other hand ipadic was created on a large corpus and thus provides disambiguation rules using contextual information. This is clearly visible when the category "other" is compared, which consists of such annotations as "symbols", or "unknownw words". The number of "other" annotations with jumandic is over two times larger than with ipadic and covers nearly 40% of the whole corpus. The detailed analysis also revealed more generic differences in word coverage of the dictionaries. Especially when it comes to abbreviations and casual modifications, some words do not appear in jumandic. For example, an interjection いや *iya* ("oh") appears in both, but its casual modification いやー *iyaa* ("ooh") appears only in ipadic. In this situation jumandic splits the word in two parts: いや and a vowel prolongation mark ー, which is annotated by jumandic as "symbol".

YACIS vs jBlogs and JENAAD: It is difficult to manually evaluate annotations on a corpus as large as YACIS[27]. However, the larger the corpus is the more statistically reliable are the observable tendencies of annotated phenomena. Therefore it is possible to evaluate the accurateness of annotations by comparing tendencies between different corpora. To verify part-of-speech tagging we compared tendencies in annotations between YACIS, jBlogs mentioned in section 2.1 and JENAAD [37]. 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. jBlogs and JENAAD were annotated with ChaSen, while YACIS with MeCab. However, as mentioned in section 4.1, ChaSen and MeCab in their default settings use the same ipadic dictionary. 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 parts-of-speech distribution is more similar between the pair YACIS(ipadic)–JENAAD (ρ = 1.0 in Spearman's rank setting correlation test) and YACIS(ipadic)–jBlogs (ρ = 0.96), than between the pairs YACIS(jumandic)–jBlogs (ρ = 0.79), YACIS(jumandic)–JENAAD (ρ = 0.85) and between both version of YACIS (ρ = 0.88).

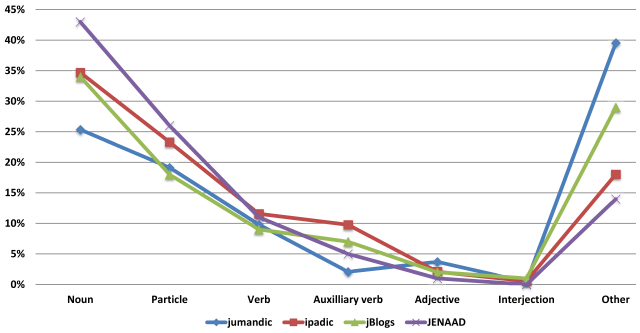

Figure 3. Graphical visualization of parts-of-speech comparison between YACIS (ipadic and jumandic annotations), Baroni&Ueyama's jBlogs and JENAAD.

Japanese vs British and Italian: As an interesting additional exercise we compared YACIS to Web corpora in different languages. In particular, we analyzed ukWaC and itWaC described in [10]. Although not all information on part-of-speech statistics is provided for those two corpora, the available information shows interesting differences between part-of-speech distribution among languages[28]. 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. Further analysis of this phenomenon could contribute to the fields of language anthropology, and philosophy of language in general.

5.2 Affective Information

Evaluation of Affective Annotations: Firstly, we needed to confirm the performance of affect analysis systems on YACIS, since the per-

[27] Having one sec. to evaluate one sentence, one evaluator would need 11.2 years to verify the whole corpus (354 mil. sentences).

[28] We do not get into a detailed discussion on differences between POS taggers for different languages, neither the discussion on whether the same POS names (like noun, verb, or adjective) represent similar concepts among different languages (see for example [26] or [22]). These two discussions, although important, are beyond the scope of this paper.

Table 7. Comparison of parts of speech distribution across corpora (with percentage).

Part of speech	YACIS-ipadic		YACIS-jumandic		jBlogs	JENAAD	ukWaC	itWaC
	percentage	(number)	percentage	(number)	(approx.)	(approx.)		
Noun	34.69%	(1,942,930,102)	25.35%	(1,419,508,028)	34%	43%	1,528,839	941,990
Particle	23.31%	(1,305,329,099)	19.14%	(1,072,116,901)	18%	26%	[not provided]	[not provided]
Verb	11.57%	(647,981,102)	9.80%	(549,048,400)	9%	11%	182,610	679,758
Auxiliary verb	9.77%	(547,166,965)	2.07%	(115,763,099)	7%	5%	[not provided]	[not provided]
Adjective	2.07%	(116,069,592)	3.70%	(207,170,917)	2%	1%	538,664	706,330
Interjection	0.56%	(31,115,929)	0.40%	(22,096,949)	<1%	<1%	[not provided]	[not provided]
Other	18.03%	(1,010,004,306)	39.55%	(2,214,892,801)	29%	14%	[not provided]	[not provided]

formance is often related to the type of test set used in evaluation. ML-Ask was positively evaluated on separate sentences and on an online forum [33]. However, it was not yet evaluated on blogs. Moreover, the version of ML-Ask supported by CAO has not been evaluated thoroughly as well. In the evaluation we used a test set created by Ptaszynski et al. [34] for the evaluation of CAO. It consists of thousand sentences randomly extracted from YACIS and manually annotated with emotion classes by 42 layperson annotators in an anonymous survey. There are 418 emotive and 582 non-emotive sentences. We compared the results on those sentences for ML-Ask, CAO (described in detail in [34]), and both systems combined. The results showing accuracy, calculated as a ratio of success to the overall number of samples, are summarized in Table 8. The performance of discrimination between emotive and non-emotive sentences of ML-Ask baseline was a high 98.8%, which is much higher than in original evaluation of ML-Ask (around 90%). This could indicate that sentences with which the system was not able to deal with appear much less frequently on Ameblo. As for CAO, it is capable of detecting the presence of emoticons in a sentence, which is partially equivalent to detecting emotive sentences in ML-Ask. The performance of CAO was also high, 97.6%. This was due to the fact that grand majority of emotive sentences contained emoticons. Finally, ML-Ask supported with CAO achieved remarkable 100% accuracy. This was a surprisingly good result, although it must be remembered that the test sample contained only 1000 sentences (less than 0.0003% of the whole corpus). Next we verified emotion class annotations on sentences. The baseline of ML-Ask achieved slightly better results (73.4%) than in its primary evaluation [33] (67% of balanced F-score with P=85.7% and R=54.7%). CAO achieved 80.2%. Interestingly, this makes CAO a better affect analysis system than ML-Ask. However, the condition is that a sentence must contain an emoticon. The best result, close to 90%, was achieved by ML-Ask supported with CAO. We also checked the results when only the dimensions of valence and activation were taken into account. ML-Ask achieved 88.6%, CAO nearly 95%. Support of CAO to ML-Ask again resulted in the best score, 97.5%.

Table 8. Evaluation results of ML-Ask, CAO and ML-Ask supported with CAO on the test set.

	emotive/ non-emotive	emotion classes	2D (valence and activation)
ML-Ask	98.8%	73.4%	88.6%
CAO	97.6%	80.2%	94.6%
ML-Ask+CAO	100.0%	89.9%	97.5%

Table 9. Statistics of emotive sentences.

# of emotive sentences	233,591,502
# of non-emotive sentence	120,408,023
ratio (emotive/non-emotive)	1.94
# of sentences containing emoteme class:	
- interjections	171,734,464
- exclamative marks	89,626,215
- emoticons	49,095,123
- endearments	12,935,510
- vulgarities	1,686,943
ratio (emoteme classes in emotive sentence)	1.39

Table 10. Emotion class annotations with percentage.

emotion class	# of sentences	%	emotion class	# of sentences	%
joy	16,728,452	31%	excitement	2,833,388	5%
dislike	10,806,765	20%	surprize	2,398,535	5%
fondness	9,861,466	19%	gloom	2,144,492	4%
fear	3,308,288	6%	anger	1,140,865	2%
relief	3,104,774	6%	shame	952,188	2%

Statistics of Affective Annotations: At first we checked the statistics of emotive and non-emotive sentences, and its determinant features (emotemes). There were nearly twice as many emotive sentences than non-emotive (ratio 1.94). This suggests that the corpus is biased in favor of emotive contents, which could be considered as a proof for the assumption that blogs make a good base for emotion related research. When it comes to statistics of each emotive feature (emoteme), the most frequent class were interjections. This includes interjections separated by MeCab (see Table 7) and included in ML-Ask database. Second frequent was the exclamative marks class, which includes punctuation marks suggesting emotive engagement (such as "!", or "??"). Third frequent emoteme class was emoticons, followed by endearments. As an interesting remark, emoteme class that was the least frequent were vulgarities. As one possible interpretation of this result we propose the following. Blogs are social space, where people describe their experiences to be read and commented by other people (friends, colleagues). The use of vulgar language could discourage potential readers from further reading, making the blog less popular. Next, we checked the statistics of emotion classes annotated on emotive sentences. The results are represented in Table 10. The most frequent emotions were joy (31%), dislike (20%) and fondness (19%), which covered over 70% of all annotations. However, it could happen that the number of expressions included in each emotion class database influenced the number of annotations (database containing many expressions has higher probability to gather more annotations). Therefore we verified if there was a correlation between the number of annotations and the number of emotive expressions in each emotion class database. The verification was based on Spearman's rank correlation test between the two sets of numbers. The test revealed no statistically significant correlation between the two types of data, with ρ=0.38.

Comparison with Other Emotion Corpora: Firstly, we compared YACIS with KNB. The KNB corpus was annotated mostly for the need of sentiment analysis and therefore does not contain any infor-

Table 11. Comparison 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 (ML-Ask)**	only	22,381,992	12,837,728	1.74
	only+ mostly	23,753,762	13,605,514	1.75

* p<.05, ** p<.01

Table 12. Comparison of number of emotive expressions appearing in three different corpora with the results of Spearman's rank correlation test.

	Minato et al.	YACIS	Nakamura
dislike	355	14,184,697	532
joy	295	22,100,500	224
fondness	205	13,817,116	197
sorrow	205	2,881,166	232
anger	160	1,564,059	199
fear	145	4,496,250	147
surprise	25	3,108,017	129
	Minato et al. and Nakamura	**Minato et al. and YACIS**	**YACIS and Nakamura**
Spearman's ρ	0.88	0.63	0.25

mation on specific emotion classes. However, it is annotated with emotion valence for different categories valence is expressed in Japanese, such as *emotional attitude* (e.g., "to feel sad about X" [NEG], "to like X" [POS]), *opinion* (e.g., "X is wonderful" [POS]), or *positive/negative event* (e.g., "X broke down" [NEG], "X was awarded" [POS]). We compared the ratios of sentences expressing positive to negative valence. The comparison was made for all KNB valence categories separately and as a sum. In our research we do not make additional sub-categorization of valence types, but used in the comparison ratios of sentences in which the expressed emotions were of only positive/negative valence and including the sentences which were mostly (in majority) positive/negative. The comparison is presented in table 11. 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). Although the numbers of compared sentences differ greatly, the fact that the ratio remains similar across the two different corpora suggests that the Japanese express in blogs more positive than negative emotions.

Next, we compared the corpus created by Minato et al. [29]. This corpus was prepared on the basis of an emotive expression dictionary. Therefore we compared its statistics not only to YACIS, but also to the emotive lexicon used in our research (see section 4.2 for details). Emotion classes used in Minato et al. differ slightly to those used in our research (YACIS and Nakamura's dictionary). For example, they use class name "hate" to describe what in YACIS is called "dislike". Moreover, they have no classes such as excitement, relief or shame. To make the comparison possible we used only the emotion classes appearing in both cases and unified all class names. The results are summarized in Table 12. There was no correlation between YACIS and Nakamura (ρ=0.25), which confirms the results calculated in previous paragraph. A medium correlation was observed between YACIS and Minato et al. (ρ=0.63). Finally, a strong correlation was observed between Minato et al. and Nakamura (ρ=0.88), which is the most interesting observation. Both Minato et al. and Nakamura are in fact dictionaries of emotive expressions. The dictionaries were collected in different times (difference of about 20 years), by people with different background (lexicographer vs. language teacher), based on different data (literature vs. conversation) assumptions and goals (creating a lexicon vs. Japanese language teaching). The only similarity is in the methodology. In both cases the dictionary authors collected expressions considered to be emotion-related. The fact that they correlate so strongly suggests that for the compared emotion classes there could be a tendency in language to create more expressions to describe some emotions rather than the others (dislike, joy and fondness are often some of the most frequent emotion classes). This phenomenon needs to be verified more thoroughly in the future.

6 CONCLUSIONS AND FUTURE WORK

In this paper we presented our research on the creation and annotation of YACIS, a large scale corpus of Japanese blogs compiled for the need of research in NLP and Emotion Processing in Text. We developed a set of tools for corpus compilation and successfully compiled the corpus from Ameblo blog service and annotated it with syntactic and affective information.

The syntactic information we annotated included tokenization, parts of speech, lemmatization, dependency structure, and named entities. The annotated corpus was compared to two other corpora in Japanese, and additionally to two corpora in different languages (British English and Italian). 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 address these differences in more detail in the future.

The affective information annotated on YACIS included emotion classes, emotive expressions, emotion valence and activation. The systems used in the annotation process include ML-Ask, a system for affect analysis of utterances and CAO, a system for affect analysis of emoticons. The evaluation on a test sample of annotations showed sufficiently high results. The comparison to other emotion corpus showed similarities in the ratio of expressions of positive to negative emotions and a high correlation between two different emotive expression dictionaries.

Although some work still needs to be done, YACIS corpus, containing over 5.6 billion words, is a valuable resource and could contribute greatly to numerous research, including research on emotions in language, sentiment and affect analysis.

YACIS corpus is meant to be used for pure scientific purposes and will not be available on sale. However, we are open to make the corpus available to other researchers after specifying applicable legal conditions and obtaining full usage agreement. In the near future we will release an additional n-gram version of the corpus to be freely accessible from the Internet without limitations and provide a demo viewable online allowing corpus querying for all types of information.

Acknowledgment

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

REFERENCES

[1] Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale and Mark Johnson. 2000. "BLLIP 1987-89 WSJ Corpus Release 1", Linguistic Data Consortium, Philadelphia, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000T43

[2] Chikara Hashimoto, Sadao Kurohashi, Daisuke Kawahara, Keiji Shinzato and Masaaki Nagata. 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.

[3] Kazuyuki Matsumoto, Yusuke Konishi, Hidemichi Sayama, Fuji Ren. 2011. "Analysis of Wakamono Kotoba Emotion Corpus and Its Application in Emotion Estimation", *International Journal of Advanced Intelligence*, Vol.3,No.1,pp.1-24.

[4] Saima Aman and Stan Szpakowicz. 2007. "Identifying Expressions of Emotion in Text". In *Proceedings of the 10th International Conference on Text, Speech, and Dialogue (TSD-2007)*, Lecture Notes in Computer Science (LNCS), Springer-Verlag.

[5] Changqin Quan and Fuji Ren. 2010. "A blog emotion corpus for emotional expression analysis in Chinese", Computer Speech & Language, Vol. 24, Issue 4, pp. 726-749.

[6] Irena Srdanovic Erjavec, Tomaz Erjavec and Adam Kilgarriff. 2008. "A web corpus and word sketches for Japanese", *Information and Media Technologies*, Vol. 3, No. 3, pp.529-551.

[7] Marco Baroni and Motoko Ueyama. 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*, www.tokuteicorpus.jp/result/pdf/2006_004.pdf

[8] Taku Kudo and Hideto Kazawa. 2009. "Japanese Web N-gram Version 1", Linguistic Data Consortium, Philadelphia, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T08

[9] Michal Ptaszynski, Rafal Rzepka and Kenji Araki. 2010. "On the Need for Context Processing in Affective Computing", In *Proceedings of Fuzzy System Symposium (FSS2010)*, Organized Session on Emotions, September 13-15.

[10] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, Eros Zanchetta. 2008. "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora", Kluwer Academic Publishers, Netherlands.

[11] Sasaki Y., Isozaki H., Taira H., Hirao T., Kazawa H., Suzuki J., Kokuryo K., Maeda E., "SAIQA : A Japanese QA System Based on a Large-Scale Corpus" [in Japanese], IPSJ SIG Notes 2001(86), pp. 77-82, 2001-09-10, Information Processing Society of Japan (IPSJ), http://ci.nii.ac.jp/naid/110002934347 (Retrieved in: 2011.11.11)

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

[13] George K. Zipf (1935) The Psychobiology of Language. Houghton-Mifflin.

[14] George K. Zipf (1949) Human Behavior and the Principle of Least Effort. Addison-Wesley.

[15] Jan Pomikálek, Pavel Rychlý and Adam Kilgarriff. 2009. "Scaling to Billion-plus Word Corpora", In *Advances in Computational Linguistics*, *Research in Computing Science*, Vol. 41, pp. 3-14.

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

[17] Vinci Liu and James R. Curran. 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.

[18] Peter D. Turney and Michael L. Littman. 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).

[19] Adam Kilgarriff, "Googleology is Bad Science", Last Words in: Computational Linguistics Volume 33, Number 1,

[20] Irena Srdanović Erjavec, Tomaž Erjavec, Adam Kilgarriff, "A web corpus and word sketches for Japanese", Information and Media Technologies 3(3), 529-551, 2008.

[21] Kilgarriff, A., Rychly, P., Smrž, P. and Tugwell, D., "The Sketch Engine", Proc. EURALEX. Lorient, France. 105-116, 2004.

[22] Jürgen Broschart. 1997. "Why Tongan does it differently: Categorial Distinctions in a Language without Nouns and Verbs." *Linguistic Typology*, Vol. 1, No. 2, pp. 123-165.

[23] Katarzyna Głowińska and Adam Przepiórkowski. 2010. "The Design of Syntactic Annotation Levels in the National Corpus of Polish", In *Proceedings of LREC 2010*.

[24] Peter Halacsy, Andras Kornai, Laszlo Nemeth, Andras Rung, Istvan Szakadat and Vikto Tron. 2004. "Creating open language resources for Hungarian". In *Proceedings of the LREC*, Lisbon, Portugal.

[25] Ichiro Hiejima. 1995. *A short dictionary of feelings and emotions in English and Japanese*, Tokyodo Shuppan.

[26] Paul J. Hopper and Sandra A. Thompson. 1985. "The Iconicity of the Universal Categories 'Noun' and 'Verbs'". In *Typological Studies in Language: Iconicity and Syntax*. John Haiman (ed.), Vol. 6, pp. 151-183, Amsterdam: John Benjamins Publishing Company.

[27] Daisuke Kawahara and Sadao Kurohashi. 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.

[28] Radoslaw Komuda, Michal Ptaszynski, Yoshio Momouchi, Rafal Rzepka, and Kenji Araki. 2010. "Machine Moral Development: Moral Reasoning Agent Based on Wisdom of Web-Crowd and Emotions", *Int. J. of Computational Linguistics Research*, Vol. 1 , Issue 3, pp. 155-163.

[29] Junko Minato, David B. Bracewell, Fuji Ren and Shingo Kuroiwa. 2006. "Statistical Analysis of a Japanese Emotion Corpus for Natural Language Processing", *LNCS* 4114.

[30] Akira Nakamura. 1993. "Kanjo hyogen jiten" [Dictionary of Emotive Expressions] (in Japanese), Tokyodo Publishing, Tokyo, 1993.

[31] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka and Kenji Araki. 2009. "A System for Affect Analysis of Utterances in Japanese Supported with Web Mining", *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, Vol. 21, No. 2, pp. 30-49 (194-213).

[32] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka and Kenji Araki. 2009. "Towards Context Aware Emotional Intelligence in Machines: Computing Contextual Appropriateness of Affective States". In *Proceedings of Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09)*, Pasadena, California, USA, pp. 1469-1474.

[33] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka and Kenji Araki. 2009. "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)*, pp. 223-228.

[34] Michal Ptaszynski, Jacek Maciejewski, Pawel Dybala, Rafal Rzepka and Kenji Araki. 2010. "CAO: Fully Automatic Emoticon Analysis System", In *Proc. of the 24th AAAI Conference on Artificial Intelligence (AAAI-10)*, pp. 1026-1032.

[35] Michal Ptaszynski, Rafal Rzepka, Kenji Araki and Yoshio Momouchi. 2012. "A Robust Ontology of Emotion Objects", In *Proceedings of The Eighteenth Annual Meeting of The Association for Natural Language Processing (NLP-2012)*, pp. 719-722.

[36] James A. Russell. 1980. "A circumplex model of affect". *J. of Personality and Social Psychology*, Vol. 39, No. 6, pp. 1161-1178.

[37] Masao Utiyama and Hitoshi Isahara. 2003. "Reliable Measures for Aligning Japanese-English News Articles and Sentences". *ACL-2003*, pp. 72-79.

[38] Janyce Wiebe, Theresa Wilson and Claire Cardie. 2005. "Annotating expressions of opinions and emotions in language". *Language Resources and Evaluation*, Vol. 39, Issue 2-3, pp. 165-210.

[39] Theresa Wilson and Janyce Wiebe. 2005. "Annotating Attributions and Private States", In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II*, pp. 53-60.

[40] Annie Zaenen and Livia Polanyi. 2006. "Contextual Valence Shifters". In *Computing Attitude and Affect in Text*, J. G. Shanahan, Y. Qu, J. Wiebe (eds.), Springer Verlag, Dordrecht, The Netherlands, pp. 1-10.

[41] T. Kudo "MeCab: Yet Another Part of Speech and Morphological analyzer", http://mecab.sourceforge.net/

[42] http://code.google.com/p/cabocha/

[43] Information Retrieval and Extraction Exercise, http://nlp.cs.nyu.edu/irex/index-e.html

[44] H. Kubota, K. Yamashita, T. Fukuhara, T. Nishida, "POC caster: Broadcasting Agent Using Conversational Representation for Internet Community" [in Japanese], Transactions of the Japanese Society for Artificial Intelligence, AI-17, pp. 313-321, 2002

[45] Okuno Yoo and Sasano Manabu, "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, 2011.

[46] Tony Berber Sardinha, José Lopes Moreira Filho, Eliane Alambert, "The Brazilian Corpus", American Association for Corpus Linguistics 2009, Edmonton, Canada, October 2009. http://corpusbrasileiro.pucsp.br

[47] Thorsten Brants, Alex Franz, "Web 1T 5-gram Version 1", Linguistic Data Consortium, Philadelphia, 2006, http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html