Language Homogeneity in the Japanese Wikipedia

Karl-André Skevik

Inferno Nettverk A/S,
Forskningsparken, Gaustadalléen 21, NO-0349 OSLO, Norway
karls@inet.no

Abstract. Wikipedia is a potentially very useful source of information, but intuitively it is difficult to have confidence in the quality of an encyclopedia that anyone can modify. One aspect of correctness is writing style, which we examine in a computer based study of the full Japanese Wikipedia. This is possible because Japanese is a language with clearly distinct writing styles using e.g., different verb forms.

We find that the writing style of the Japanese Wikipedia is largely consistent with the style guidelines for the project. Exceptions appear to occur primarily in articles with a small number of changes and editors.

Keywords: wikipedia, japanese, nlp

1 Introduction

Wikipedia[1] is a comprehensive Internet based encyclopedia, which is unusual in that the articles are largely written and maintained by volunteers, many of whom are anonymous. The permissive approach used by Wikipedia does however make it natural to question the correctness of the content, and the possibility of achieving a consistent writing style. An encyclopedia is expected to have a formal writing style consistently applied across all articles and in this paper, we examine the writing style of the entire Japanese Wikipedia using automated analysis. This approach is motivated by the presence in Japanese of very distinct writing styles that can serve to signal different degrees of formality and politeness. The difference in language usage is not limited only to word choice but includes different forms for sentence-final elements. By examining all sentences in Wikipedia, we attempt to identify any inconsistencies in writing style.

Section 2 of this article lists related work, and Section 3 discusses Japanese writing styles. In Section 4, our analysis procedure is described, with the classification results presented in Section 5. Some of the Wikipedia articles that did not follow the style guidelines are analyzed in Section 6 and our conclusions are presented in Section 7.

2 Related Work

Most of the existing research on Wikipedia falls into one of four categories; research that looks at the procedures for avoiding incorrectness and detecting errors in articles, research that attempts to evaluate content correctness, research that looks at completeness, and research that proposes ways in which the article quality can be automatically analyzed.

2.1 Process evaluation

The English Wikipedia community has according to Stvilia et al. (2008) developed extensive processes to achieve article quality. Instead of error prevention, Wikipedia makes use of techniques that allow problems to be fixed quickly when they occur.

[1] http://www.wikipedia.org

Lorenzen (2006) examined the contents of a Wikipedia page used to report potential vandalism over a period of several months and found that a significant amount of resources are spent on addressing this issue. Detecting this in less popular articles can be difficult, but controversial and frequently vandalized pages are likely to be quickly corrected. Viégas et al. (2004) find that vandalized pages are fixed in a median time of 1.7 minutes.

2.2 Correctness

Emigh and Herring (2005) have looked at the extent to which Wikipedia, and one other similar project, produce work that is similar or dissimilar to existing print encyclopedias. This was done by looking at the formality of the language in use. Based on the content of 15 articles, the degree of formality was quantified by separately counting the occurrence of word usage typical of both formal and informal English language usage genres. For the informal genres this included contractions and personal pronouns, while noun formative suffixes was used to measure the degree of formality. In addition, the average word length and number of words in a sentence was calculated. The results for these articles show that the level of formality is close to that of the print encyclopedia, while the content of the discussion pages is far more informal. This paper makes a similar survey of the Japanese Wikipedia, looking at the formality and consistency of language usage, but rather than manually examining a limited number of articles, we study the entire Japanese Wikipedia.

Nielsen (2007) examined the use of citations of scientific journals in the English Wikipedia, and found these to have citation usage correlating with that of scientific journals. An expert-led peer-review performed by Nature compared Wikipedia to the Encyclopedia Britannica, by examining 42 articles (Giles, 2005). Errors were found in both encyclopedias, with an average of three errors found in the science articles of the Encyclopedia Britannica, compared to an average of four in Wikipedia.

Wilkinson and Huberman (2007) have looked at the correlation between article quality and factors such as the number of edits and distinct editors for an article. The authors determined that a high number of editors and edits is indicative of high article quality.

2.3 Completeness

Devgan et al. (2007) study the accuracy of medical information in Wikipedia, using two independent reviewers to examine a selection of articles on common medical procedures. The authors found that, though not complete, the Wikipedia entries were accurate.

The completeness of drug information is studied by Clauson et al. (2008), by comparing Wikipedia to a specialized drug database. Wikipedia was found to be less complete. A similar study for medical informatics done by Altmann (2005) also found many basic concepts to be missing.

2.4 Automated process proposals

HU et al. (2007) propose a way of automatically calculating article quality rankings based on the retention of content changes and additions made by editors. Potthast et al. (2008) suggest a way of automatically detecting vandalism by looking at characteristics of typical changes made by vandals. Adler and de Alfaro (2007) propose a reputation system for Wikipedia editors, based on the degree to which changes made by editors remain in Wikipedia. Similar approaches for automatically analyzing the quality and trustworthiness of article content is proposed by Dondio and Barrett (2007), and McGuinness et al. (2006).

3 Japanese

In Japanese, there are several different types of expression alternatives with the same meaning, but with differences in politeness, roughness, formality, elegance, and vulgarity (Wetzel, 2004,

p. 39). Proper usage is determined by place and time, whether the context is written or spoken, interpersonal relationships, and psychological factors such as the intent of the speaker.

One way of classifying these alternatives is through the three primary honorific processes in Japanese. Given a speaker, an addressee, and a referent, where the referent can be the speaker, the addressee, or a third party, Shibatani (1991, p. 375-376) describes Japanese as having honorific processes along two independent axes: the speaker-addressee axis and the speaker-referent axis. The first of these is also referred to as *addressee controlled honorifics*, while the speaker-referent axis consists of so-called *subject honorifics* and *object honorifics*. These three terms roughly correspond to the Japanese terms *teineigo* (polite language), *sonkeigo* (respect language), and *kenjougo* (humility language). Most of these are however primarily relevant in spoken language, or writing with a known recipient; there are other expectations when there is no specific reader (Shibatani, 1991, p. 360), such as is the case with an encyclopedia.

For the Japanese Wikipedia, there is especially one page that covers writing style[2]. The guidelines are, according to the page, approved by many users but are not official policy. The general goal of the guidelines is to achieve consistency and a style which is close to that of printed text. For the encyclopedia articles the recommended practice is use of plain forms, including the *de aru* and *da* copula forms.

To verify that the expected writing style is in use it then primarily becomes necessary to examine *teineigo* usage; whether the forms are plain or polite. For example, for the verb “to write”, the plain form is *kaku*, and the polite form *kakimasu*. For the copula, the plain form is *da*, and the polite form *desu*. The *de aru* plain copula form is a more formal variant typically used in newspapers and scholarly articles.

Japanese permits some reordering of sentence elements, but is basically a subject-object-verb (SOV) language that requires the verbal element to come last. It is primarily also for the sentence-final verb that the choice between plain and polite forms exist. There are other relevant aspects of writing style that are not found at the end of a sentence, but these are more difficult to classify due to ambiguity and the possibility of ellipsis. For example, the *-rare* suffix is homophonous with the suffix for the passive, potential, and spontaneous forms (Shibatani, 1991, p. 375). A way of analyzing conversations in order to identify the topic, object, or subject in sentences where these elements are not explicitly stated is presented by Yoshimoto (1988), but there are cases when this information cannot easily be deduced, a problem also noted by Shirado et al. (2006).

In this paper, we focus mainly on the form of the last sentences element, which is easily identified and gives a meaningful result for each sentence. We used *MeCab*[3] for sentence analysis.

4 Analysis procedure

A Wikipedia XML snapshot file is taken as input and preprocessed to remove XML markup and other non-text elements. The result is a sequence of Japanese sentences, which are given as input to MeCab for morphological analysis. The final part of each sentence, as identified by MeCab, is then analyzed with a tool we wrote in *perl* to identify various sentence characteristics.

Wikipedia contains three page types which we analyze. In addition to the article pages with the type of content found in more traditional encyclopedias, there are *user pages* where the Wikipedia editors introduce themselves, and per-article *discussion pages* used by editors to discuss changes and resolve disagreements. The content of these pages are in other words written by the same people that contribute to the article pages.

We obtained the Wikipedia article data from a downloaded snapshot marked with the date July

[2] The title of this article in the Japanese Wikipedia is “Wikipedia:表記ガイド” (publishing guide). We examined the version of the page last modified May 1, 2009.

[3] `http://mecab.sourceforge.net/`

24, 2008[4]. The contents of the discussion and user pages were obtained from the file *jawiki-20080724-pages-meta-current.xml*.

Lines that end with a Japanese full stop character, exclamation mark, or question mark are considered to be valid sentences. The remaining lines typically contain text like poems, incomplete sentences, or difficult to detect markup, and are removed before processing with MeCab. After preprocessing, the encyclopedia data contains $6,510,554$ sentences, the discussion page data $697,278$ sentences, and the user page data $1,072,337$ sentences.

For the encyclopedia data, 13.6% of the articles have no valid lines, while as many as 75.8% of the articles have no invalid lines. In total, 94.7% of the articles have no more than one invalid line. For the user pages, 28.1% of the articles have no valid lines, and 81.2% have no invalid lines. The number of articles with at most one invalid line is 95.6%. Of the discussion pages, only 2.9% of the pages have no valid lines, and 72.6% have no improper lines. At most one invalid line is found in 90.3% of the article pages.

Overall, the extracted sentences should give a representative overview of the language used in the Japanese Wikipedia. The content extraction is fairly efficient, if not perfect.

5 Initial classifier distribution

Some of the most frequently occurring characteristics in the sentences in the encyclopedia articles are shown in Table 1(a). This list contains a subset of the most frequently occurring sentence categories[5]. The *(unclassified)* entry corresponds to sentence patterns not supported by our tool. We added support for the most frequently occurring entries, leaving 1.01% of all sentences in the encyclopedia data as unclassified. The characteristics generally name the class of the last word in the sentence and is followed by the name of a Japanese particle or auxiliary verb. For example, sentences classified as *noun+da* end with a noun and the *da* copula form. The *(plain)* value indicates usage of plain verb forms, while *dneg (plain)* corresponds to the negative form *dehanai*.

(a) Encyclopedia articles

Characteristic	Pct.	Cum. pct.
verb (plain)	60.90	60.90
noun+none (plain)	21.60	82.50
dearu (plain)	10.77	93.27
iadj+none (plain)	2.21	95.49
(unclassified)	1.01	96.51
noun+da (plain)	0.64	97.15
shimau (plain)	0.24	97.40
noun+dneg (plain)	0.23	97.63
verb+masu (polite)	0.12	97.76
dearou (plain)	0.06	97.83
nadj+dneg (plain)	0.04	97.87
nadj+da (plain)	0.04	97.92
verb+yasui (plain)	0.04	97.96
noun+desu (polite)	0.03	98.00
verb+te+kudasai (-)	0.02	98.02
noun+darou (-)	0.01	98.04
verb+nikui (plain)	0.01	98.05

(b) Discussion pages

Characteristic	Pct.	Cum. pct.
verb+masu (polite)	44.50	44.50
noun+desu (polite)	12.47	56.97
(unclassified)	6.79	63.76
verb (plain)	6.24	70.01
noun+none (plain)	5.52	75.54
noun+deshou (-)	4.69	80.23
verb+deshou (-)	2.68	82.91
verb+te+kudasai (-)	1.95	84.86
nadj+desu (polite)	1.61	86.47
iadj+desu (polite)	0.92	87.40
noun+dneg+masu (polite)	0.84	88.24
dearu (plain)	0.82	89.06
noun+kudasai (-)	0.66	89.73
iadj+none (plain)	0.65	90.38
noun+da (plain)	0.37	90.76
te+masu (polite)	0.34	91.10

Table 1: Sentence characteristics (i)

The majority (60.90%) of all sentences in the encyclopedia articles end in a plain verb, which is consistent with Japanese as a SOV language. The *de aru* copula form is also used quite extensively,

[4] The file *jawiki-20080724-pages-articles.xml*, containing a snapshot of the Japanese Wikipedia, was obtained from: `http://download.wikimedia.org/jawiki/`

[5] The cumulative percentages only includes the values actually listed.

(a) User pages

Characteristic	Pct.	Cum. pct.
verb+masu (polite)	48.77	48.77
"youkoso" (-)	6.28	55.06
noun+desu (polite)	4.79	59.85
"hajimemashite" (-)	4.30	64.16
verb+te+kudasai (-)	4.23	68.39
noun+none (plain)	4.07	72.47
noun+kudasai (-)	3.72	76.19
verb (plain)	3.40	79.60
"konnichiha" (-)	3.29	82.89
(unclassified)	3.28	86.18
nadj+desu (polite)	2.05	88.24
o+verbstem+kudasai (-)	1.74	89.99
noun+desyou (-)	1.00	91.00
noun+dneg+masu (polite)	0.70	91.71
dearu (plain)	0.64	92.35
iadj+desu (polite)	0.38	92.73
"arigatou gozaimashita" (-)	0.35	93.08
verb-te (plain)	0.34	93.43
iadj+none (plain)	0.25	93.68
noun+itashimashita (polite)	0.19	93.87
noun+da (plain)	0.14	94.02

Table 2: Sentence characteristics (ii)

being the sentence-final element in 10.77% of all sentences, while the *dearou* form occurs in 0.06% of all sentences. Usage of plain verb forms and the *de aru* copula form is consistent with the proper writing style for Wikipedia (see Section 3). There are however 8033 sentences that use polite verbs forms (0.12%), and we study these in more detail below.

For nouns, ellipsis of the element following it is most frequent, occurring in 21.60% of the sentences. We currently do not try to automatically determine whether sentences have ellipsis of the sentence-final verb or of the copula, but only 0.64% of the sentences have the *da* copula form after a noun, compared with 10.77% for the *dearu form.* The number of sentences with the *da* copula form is so low that it is possible that some of the sentences ending with a noun have ellipsis of the copula, which would be contrary to what the Wikipedia style guide recommends (use of *de aru* and *da*).

A similar overview for the discussion and user pages is given in Table 1(b) and Table 2(a), respectively. The *te+masu* entry indicates use of the shorter *te-masu* form instead of *te-imasu.* Frequently occurring expressions and interjections are shown between quote marks. These occur often in the user pages, which obviously have the viewpoint of the writer: *youkoso* (welcome), *hajimemashite* (nice to meet you), and *konnichiha* (hello), are expressions that one might expect to find in personal pages.

The writing style of the discussion and user pages is not as consistent as for the encyclopedia articles. In the discussion pages, 44.50% of the verbs use the polite *masu* forms[6], but 6.24% use plain forms, compared to only 0.12% using the non-dominant form in the encyclopedia data. After nouns, the *desu* copula form is dominant, at 12.47%, but as many as 5.52% have ellipsis of the copula or verb (the *da* copula form is used in only 0.37% of all the sentences). A similar variation is seen in the user page data. Polite verb forms are dominant, at 48.77%, but 3.40% use the informal style. Ellipsis of the copula or verb after nouns occurs in 4.07% of all sentences, almost as frequently as the *desu* form, at 4.79%.

The tables show that the writing style of the encyclopedia content is clearly distinct from that

[6] This does not include the *te+masu* verbs, which come in addition.

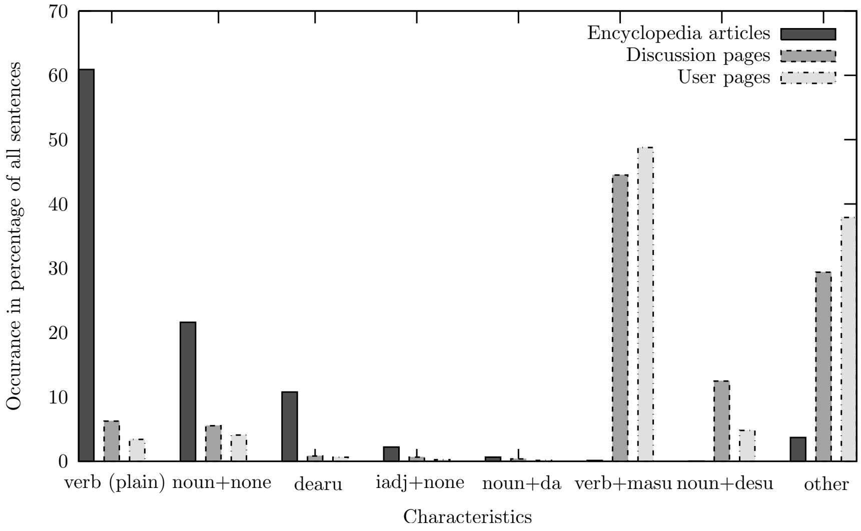

Figure 1: Initial classifier distribution

of the discussion and user pages. A comparison of the three page types is shown in Figure 1, which lists the characteristics that occur most frequently in the encyclopedia articles. Especially the difference between the frequency of plain and polite *masu* verb forms can be clearly seen. The writing style of the encyclopedia pages is also more consistent, with only a small number of characteristics needed to describe most of the sentences.

Emigh and Herring (2005) found that the language in the discussion pages of the English Wikipedia was less formal than that of the encyclopedia articles. For the Japanese Wikipedia, we can see that more polite forms are used in the user and discussion pages. However, the difference between the encyclopedia articles and the other two page types is that the first has no specified reader, while the content of the other pages is more personal and perhaps thought as meant for other editors. This difference is possibly what determines the writing style; the use of plain forms in the encyclopedia articles and polite forms in the other pages does not imply that the level of formality is the opposite of that of the English Wikipedia.

6 Style variations

The encyclopedia articles mostly use plain forms, but we have classified 0.12% of the sentences as having *masu* forms, and 0.03% as using the *desu* copula form after a noun. To determine if these sentences represent examples of the guidelines not being followed, we wrote a simple tool that lists the relevant sentences and the name of the articles they occur in. We then manually examined the output.

There were some deficiencies in the preprocessing stage used to remove unwanted content. Improving the preprocessing stage would address some of this, but more problematic is the fact that in some articles entire sections contain text written only in polite forms. Articles with text primarily taken from other sources, such as letters or story summaries for books, in some cases contain only polite forms. Removing this type of text would be difficult without some form of markup to signal that polite forms would be proper.

Finally, there are the actual cases of incorrect writing style. We manually verified 25 instances of this occurring[7]. We did not examine the entire list of possible style errors because even after ignoring the obvious instances of sentences that should have been removed during preprocessing, the list contained several thousand entries. The impression given by an examination of a subset of

[7] In some cases the errors were found to have been corrected in more recent article versions.

the list is that a large part of the sentences that use polite forms do not represent examples of inconsistent language usage, but text that either should ideally have been removed during preprocessing, or text where an editor might argue that polite forms are proper.

According to Wilkinson and Huberman (2007), errors are primarily found in articles that have few editors and few edits, and to a certain extent our results fulfill these expectations. As many as 16 of the 25 errors occur in articles with at most 17 revisions, and no more than 5 editors. As noted above, this is only a subset of the possible errors identified, but the relatively small number of lines using polite forms[8] show that the total number of errors cannot be high.

7 Conclusions

We have examined $6,510,554$ sentences from the encyclopedia articles in the Japanese Wikipedia, and for characteristics such as usage of plain or polite forms, the writing style appears to be quite consistent, despite the open nature of Wikipedia. In addition to the encyclopedia articles, we examined discussion and user pages, which are also written by Wikipedia editors, but where the same expectations for writing style do not exist. The consistency of the writing style of the encyclopedia articles is emphasized by the higher degree of variation in the discussion and user pages. This is fairly impressive, considering the collaborative way in which the articles are written.

To the extent that there is a discrepancy, it is with use of the *da* copula form, which occurs to a much lesser degree than what is possibly ellipsis of the copula. One possible way to clarify this might be with the *Japanese Web N-gram Version 1* data set (Kudo and Kazawa, 2009). For example, for nouns that are never or only rarely followed by a verb, ellipsis following the noun in a sentence final position is most likely to be ellipsis of the copula.

Our results also show the feasibility of using automated sentence analysis to verify the consistency of writing style in large distributed projects such as Wikipedia. At least for languages such as Japanese, where the writing style is relatively easy to determine, editors can use this approach as an aid to identify articles that might need to be corrected.

References

Adler, B. T. and de Alfaro, L. 2007. A content-driven reputation system for the Wikipedia. In *WWW '07: Proceedings of the 16th International World Wide Web Conference*.

Altmann, U. 2005. Representation of medical informatics in the Wikipedia and its perspectives. In *MIE '05: Proceedings of the 19th International Congress of the European Federation for Medical Informatics, Connecting Medical Informatics and Bio-Informatics*.

Clauson, K. A., Polen, H. H., Boulos, M. N. K., and Dzenowagis, J. H. 2008. Scope, completeness, and accuracy of drug information in Wikipedia. *The Annals of Pharmacotherapy*, 42(12).

Den, Y., Nakamura, J., Ogiso, T., and Ogura, H. 2008. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *LREC '08: Proceedings of the 6th International Language Resources and Evaluation*.

Devgan, L., Powe, N., Blakey, B., and Makary, M. 2007. Wiki-surgery? Internal validity of Wikipedia as a medical and surgical reference. *Journal of the American College of Surgeons*, 205(3).

Dondio, P. and Barrett, S. 2007. Computational trust in web content quality: A comparative evaluation on the Wikipedia project. *Informatica*, 31.

[8] After excluding articles and text that it can be easily determined should have been removed during the preprocessing step, there were in total 3550 lines in 1170 articles that used polite forms, which is only 0.054% of all the extracted lines. Only a subset of these lines are actual errors.

Emigh, W. G. and Herring, S. C. 2005. Collaborative authoring on the web: A genre analysis of online encyclopedias. In *HICSS-38 '05: Proceedings of the 38th Hawaii International Conference on System Sciences*.

Giles, J. 2005. Internet encyclopedias go head to head. *Nature*, 438.

HU, M., Lim, E.-P., Sun, A., Lauw, H. W., and Vuong, B.-Q. 2007. Measuring article quality in Wikipedia: Models and evaluation. In *CIKM '07: Proceedings of the 16th ACM Conference on information and knowledge management*.

Kudo, T. and Kazawa, H. 2009. Japanese Web N-Gram Version 1. Linguistic Data Consortium, Philadelphia.

Lorenzen, M. 2006. Vandals, administrators, and sockpuppets, Oh My! An ethnographic study of Wikipedia's handling of problem behavior. *MLA Forum*, 5(2).

McGuinness, D. L., Zeng, H., da Silva, P. P., Ding, L., Narayanan, D., and Bhaowal, M. 2006. Investigations into trust for collaborative information repositories: A Wikipedia case study. In *MTW '06: Proceedings of the WWW '06 Workshop on Models of Trust for the Web*.

Nielsen, F. Å. 2007. Scientific citations in Wikipedia. *First Monday*, 12(8).

Potthast, M., Stein, B., and Gerling, R. 2008. Automatic vandalism detection in Wikipedia. In *ECIR '08: Proceedings of the 30th European Conference on IR Research, Advances in Information Retrieval*.

Shibatani, M. 1991. *The Languages of Japan*. Cambridge Language Surveys.

Shirado, T., Marumoto, S., Murata, M., Uchimoto, K., and Isahara, H. 2006. A system to indicate honorific misuse in spoken Japanese. In *ICCPOL '06: Proceedings of the 21st International Conference of Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*, volume 4285, pages 403–413.

Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. 2008. Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology (JASIST)*, 59(6):983–1001.

Viégas, F. B., Wattenberg, M., and Dave, K. 2004. Studying cooperation and conflict between authors with history flow visualizations. In *CHI '04: Proceedings of the 2004 Conference on Human Factors in Computing Systems*, pages 575–582.

Voss, J. 2005. Measuring Wikipedia. In *ISSI '05: Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*.

Wetzel, P. 2004. *Keigo in Modern Japan*. University of Hawaii Press.

Wilkinson, D. M. and Huberman, B. A. 2007. Cooperation and quality in wikipedia. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*.

Yoshimoto, K. 1988. Identifying zero pronouns in Japanese dialogue. In *COLING '88: Proceedings of the 12th conference on Computational linguistics*.