Studia Linguistica
Universitatis Iagellonicae Cracoviensis 124 (2007)

ANNA TERESZKIEWICZ
Instytut Filologii Angielskiej

LANGUAGE MISTAKES AND STYLISTIC INCONSISTENCY IN THE ARTICLES OF ENGLISH AND POLISH WIKIPEDIA

Wikipedia, the free encyclopedia as it is referred to, is now the largest multilingual encyclopedia available online. Its English edition covers nearly 1,660,000 articles. The Polish edition is smaller, containing 356,600 posts. The most important characteristics of Wikipedia is that it is composed collaboratively by the users of the web. This means that any person having access to the web can edit an article without being subject to a prior editorial review. Thus, compared to a standard encyclopedia, composed by a board of editors, who are specialists in a given field of knowledge, Wikipedia is created by a large, diversified and, to a considerable extent, anonymous body of web-users who do not necessarily specialize in a chosen area of study. As indicated by Wikipedia itself, the authors are mainly students, hobbyists and volunteers who appreciate the feeling of creating a source of knowledge for others.[1]

Due to the above mentioned characteristics, Wikipedia has became the subject of various disciplinary approaches, e.g. a sociological approach, analysing the phenomenon of Wikipedia community (Viegas 2004), or a computational approach, focusing on the wiki system upon which the encyclopedia is based (Lih 2004). From a linguistic point of view, it is interesting to study how language is used within this particular setting, and to what extent the features of the system exert impact on the stylistic sphere of the encyclopedia. It is hypothesized that such openness of Wikipedia and lack of strict editorial control may lead to linguistic flaws of the entries and to occurrence of grammatical and stylistic errors (Herring 2005, Tereszkiewicz 2006). The previous study of the Polish edition of Wikipedia (Tereszkiewicz 2006) proved that the encyclopedia is characterized by linguistic heterogenity and stylistic variance.

This study concentrates on both issues of Wikipedia, English and Polish, with the focus on analyzing language mistakes and stylistic inconsistencies. The main target of the study is to extract the most frequent types of language errors,

[1] Cf. information on Wikipedia: http://en.wikipedia.org/wiki/Wikipedia:Community_Portal

grammatical and stylistic mistakes[2] occurring in the articles of the two editions of Wikipedia.

The analysis covered the sample of 200 articles edited in English and Polish Wikipedia, 100 articles each. The entries were chosen randomly, using the service "random article" provided by Wikipedia. The entries concern a wide spectrum of issues: from scientific texts, through geography and sports to humanities.

The language of an entry to an encyclopedia should be characterized by formality, objectiveness, conciseness and clarity (Crystal and Davy 1969; Wilkoń 1987). Writing an entry to an encyclopedia demands from an author a degree of linguistic awareness, linguistic skill and the ability to express in a formal register. Thus, most important is the choice of an appropriate linguistic means of expression. This choice concerns lexical, grammatical and stylistic elements. A linguistic mistake is, in this case, the choice of an inappropriate variant of expression – lexical, grammatical or stylistic. Inappropriate, in this context, implies used not in accordance with the norm (Swales 1990; Gajda 1993).

The analysis of the entries to Wikipedia has illustrated a degree of stylistic syncretism. While a large percentage of articles is written in a highly formal register, meeting the standards of an encyclopedia entry, other articles display considerable stylistic diversity in the choice of words and expressions. It is common that in one entry there appear forms belonging to different registers. Apart from stylistic inconsistencies, what differentiates the entries in Wikipedia from a standard encyclopedia is the presence of evaluative expressions. What follows is the description of the most frequent types of language mistakes and stylistic inconsistencies distinguished in the English and Polish articles.

1. Lexical mistakes

1.1. Colloquial vocabulary and expressions

Authors of articles in English and Polish Wikipedia frequently resort to the use of colloquial vocabulary. There is, however, a difference in the frequency of occurrence of such expressions. In the English Wikipedia, colloquialisms appeared in 23% of the analyzed articles, whereas in the Polish edition, this frequency was higher, amounting to 37%.

Due to the presence of such means of expression, particular entries in both editions do not have the characteristics of formality normally attributed to an encyclopedia.

Colloquial means of expression to a large extent involve the use of phrasal verbs (11%), e.g. *come upon*, *break off* or *deal with*, which could be exchanged with more formal counterparts such as *encounter* or *terminate* or *concern*.

The analysis also revealed frequent occurrence of idiomatic expressions (4%), e.g. *he took it for granted*; *he had no clue*; *who shot to fame during the*

[2] The terms *errors* and *mistakes* are used interchangeably in this article (unlike in methodology studies).

1980s; *their contribution to the Kalenjin lexicon is worth nothing*; *the film got into trouble with film censors*; *...is making a comeback in (...)*; *Thomas Cook toyed with the idea of buying them out of bankruptcy.*

In the analyzed entries, the following colloquial expressions were encountered (8%): *there came the debate*; *he had no clue what to make his next movie about*; *Brocail was involved in an ugly incident at the McAfee Coliseum*; *rumors are going that*; *there is a wave of reviving or beefing up of (re)legalized.* It may be assumed that authors do not strive to achieve a high level of formality of expression, resorting to common, every-day vocabulary.

As far as the Polish material is concerned, the following instances of colloquialisms were distinguished: idioms (7%): *co prowadzi na manowce*; *najbardziej prymitywnym systemem jaki sobie można wyobrazić*; *Włochem nie był w żadnej mierze*; *jej rumieniec przywodzi na myśl*; *szło mu nieźle*; informal vocabulary (8.5%): *dość*, *niezbyt*, *kiepsko*, *zbytnio*, *po prostu*; informal means of expression (18.5%): *skoczek lubi skocznie mamucie*; *niedługo potem odezwała się kontuzja kolana*; *podczas gdy młodszy musiał się zadowolić Rybnikiem*; *oni za wszelką cenę nie chcą przyjąć do wiadomości, że...*; *śmiertelnie się przejmują.* As noted before (Tereszkiewicz 2006) informal terms have an expressive function.

1.2. Evaluative expressions

In spite of the "neutral point of view policy" advocating impersonality and objectivity in presenting information, a number of articles in Wikipedia are characterised by subjectivity and personal commentaries. Authors of the entries do not refrain from the presentation of evaluative judgments, usage of emotive vocabulary and terms expressing subjective opinions. Evaluative judgments are signalled by a wide range of verbal items which interpret the ideational content for the reader and express a value judgment about it.

In English such expressions appeared in 10.4% of the articles and included the following examples: *Amazingly, they are able to easily understand all the other dialects of the Kalenjin*; *Furthermore, it is very important to understand that breathing pure oxygen…*; *Surprisingly, the ill-organized rebels managed to defeat the inadequate and inefficient Imperial forces*; *It is striking how the (mainly informal) terminology is usually determined by the punisher's point of view*; *some curious devices as produced for U.S. fraternity initiations*; *... an obvious joke of it being put on instead of South Park*; *from the above table, clearly there is a problem with the above theory*; *although there can be no doubt that Maya society and tradition has undergone substantial change ...*; *heretical in no small part due to Jesuit efforts.*

Also, there is a high frequency of the use of hedges (7%): *actually*; *undoubtedly*; *obviously*; *almost certainly*; *however*; *of course.*

As far as Polish entries are concerned, evaluative expressions occurred in nearly 18% of the articles: *hipoteza dość dobrze tłumaczy, że…*; *dla pięciu barw dowód jest stosunkowo prosty*; *należy zwrócić uwagę na fakt*; *to cecha polegająca, mówiąc najprościej, na bezpośrednim truciu ludzi*; *zdumiewające jest to, że dywan Sierpińskiego jest krzywą według jednej z jej definicji!*; *ta oczywista ob-*

serwacja nie może mieć poważnych zastosowań; *co ciekawe*; *łatwo zauważyć*; *warto pamiętać, że skoro nie żądamy od aksjomatu prawdziwości*; *to także nie jest prawdą*.

The function of evaluative language can be both ideational and interpersonal. Many of the evaluative words enumerated above can simultaneously express the author's value judgment and influence the reader's opinion. In using the above mentioned expressions, authors very often suggest a possible interpretation of the information presented. Such expressions show lack of emotional distance towards the presented contents. Frequently, authors take advantage of the freedom of expression offered by Wikipedia to present their own opinions and comments on particular subjects.

1.3. Erroneous collocations and pleonasms

While colloquial expressions were frequent in both editions of Wikipedia, lexical mistakes were found only in the Polish encyclopedia and appeared in nearly 10% of the analyzed articles. This type of mistake encompassed the use of erroneous collocations, recurrence of redundant elements as well as the use of pleonasms *to pogląd związany z wzajemnymi powiązaniami*; *tak iż poseł rosyjski uważał iż*; *teoria z założenia zakłada*; *udało się dokonać sprawdzenia*; *powoduje to, że wiązanie to*; *założycielem i twórcą założeń był*; *może się przydawać jako metoda*; *zaczął budować swoją karierę*; *występy zarobiły mu entuzjastyczne recenzje*; *kłopoty zdają się kochać Dutta*; *współczesna nauka tu też się chyba myliła.*

2. Grammatical mistakes

Lack of concord between clauses constitutes a frequently occurring grammatical mistake in the English articles. The most repeated examples include the lack of tense (4%) and person (3.6%) concordance, e.g.: *man was made unsuccessfully out of mud and then wood before being made out of maize and being assigned tasks which praised the gods. Then the legendary hero twins, Hunahpu and Ixbalanque descend into the underworld, perish, and are eventually miraculously reborn*; *it is known to have first existed in Mesopotamia and China and was invented sometime between 1000 BC and 500 BC.*

As well, authors of the English entries frequently interchange pronouns, which introduces vagueness and lack of clarity (2.6%), e.g.: *to get to the next landscape, the player must open all the Oxyds on the current landscape, which is done by touching them, but they will only stay open if you touch Oxyds of the same color in sequence, so one must not change it.*

In Polish the grammatical mistake of the lack of concord is less frequent. In the analyzed material, there occurred only 4 instances of such a mistake, e.g.: *to wytłumaczenie wyjaśnia także, dlaczego kwazary były o wiele powszechniejsze we wczesnym wszechświecie – produkcja tak wielkich energii kończy się, kiedy czarna dziura zje wszystkich kosmiczne śmieci wokół siebie.*

A particularly frequent grammatical mistake observed in the English Wikipedia, distinguished in nearly 6% of the articles, concerns the use of the conjunction *and*. In many cases, this conjunction is used to link information and data into one hypotactic sentence. Use of a different conjunction, or disjunction (e.g. *but*), would contribute to greater clarity and meaningfulness in the sentence, for example: *there is no mention of Babrius in ancient writers before the beginning of the 3rd century AD, and his language and style seem to show that he belonged to that period*; *Huracan, or the Heart of Heaven, also existed and is given less personification.* As used in these examples, *and* seems to impart the oral nature to the entries. In oral discourse, namely, *and* fulfills a variety of functions, such as adding a comment, pointing to contrast or chronological sequence (Quirk 1972). All these functions are visible in the entries to the encyclopedia.

A frequent grammatical mistake observed in nearly 19% of the Polish articles is the lack of a proper conjunction of clauses. Very often authors use long, hypotactic sentences, the division of which would contribute to a greater clarity of the presented contents, e.g.: *te morfologiczne właściwości nie pokrywają się z własnościami funkcjonalnymi, tzn. na podstawie morfologii danego limfocytu nie można ustalić jego funkcji, chociaż np. wiadomo, że 70% komórek NK wykazuje morfologię LGL i to właśnie one zwykle są spotykane w tej postaci, pewności jednak nigdy nie ma.*

Similarly, due to an improper or unnecessary conjunction of clauses containing different meanings, the sentences sound inappropriate and stylistically erroneous, e.g.: *Brocail attended Lamar High School in Lamar, Colorado and was a good student and won All-State honors in football, basketball, and baseball*; *he is considered to be a superlative composer both by his contemporaries and by modern scholars, however his surviving output is small, and he died young*. In Polish, the examples include the following: *lubi skocznie mamucie – na nich spisuje się dużo lepiej niż na normalnych i dużych, uzyskał imponujący rekord życiowy 221.5 m w Planicy*; *jego ojciec jest zawodowym pilotem, a Alan Alborn także posiada licencję pilota*; *poszczególne prądy łączyły się i przenikały, ze względu na płynność i ewolucję poszczególnych działaczy – wszelkie podziały są elastyczne.*

Many grammatical mistakes result from a lack of conscientious editing. Many of them involve a missing sentential element, use of an incorrect grammatical form or a wrong preposition, for instance in English (2%): *here he worked under the studio of and was involved in the studio's large projects*; *the first half to the 17th century*; *themself*; *developed beginning in 2004*; *it is expected it to fall in chart positions or even disappear.* In Polish (2.4%): *nauka dotyczący Boga*; *rozpuszczalniki dzieli się je na dwie grupy*; *nie bez znaczenie*; *nie leczone zależności.*

3. Punctuation mistakes

Though punctuation and spelling mistakes are few in Wikipedia (in English 2%, in Polish 3.4%), in comparison with the frequency of the above mentioned mistakes, it is worth pointing to them. Mostly, they involve lack of punctuation marks, e.g. lack of a comma, final dot, capital letters. In most cases, these mistakes take place because of mere negligence and lack of careful editing. We encountered the following spelling and punctuation mistakes: in English: *to don painted clothe, Various religious movement*; *employs small children who reportedly suffer repeated systemic human rights abuses*; *one example would be if a person owns a house but wants to buy a second house*; *the dike however was not finished yet.* In Polish: *na ogół aby wiązanie się wytworzyło*; *wiązanie formalnie rzecz biorąc*; *czym są bardziej przestrzennie zbudowane tym ich czas życia się wydłuża*; *reagują same ze sobą na skutek czego*; *oznacza zrobienie porządków, posprzątanie*; *tu poznał się ze swą żoną Ann. Z którą uciekł na południe*; *początkowo pracował z ojcem ale odszedł.*

4. Stylistic syncretism

A particularly interesting and frequently occurring feature of Wikipedia entries is a specific stylistic eclectism. Among the analyzed articles there were many instances of a peculiar stylistic syncretism, where formal expressions interchanged with colloquial terms. Such interchangeability of different registers may signify lack of ability or awareness of the need to distinguish between the different registers that should be used in different contexts. To the examples of this intertextuality in English belong the following: *Orchestration is the study or practice of writing music for orchestra (or, more loosely, for any musical ensemble) or of adapting for orchestra music composed for another medium. It only gradually over the course of music history came to be regarded as a compositional art in itself*; *The designs were modified to include more human-like features, including a mannequin-like face. Other issues included the conditions that the robots and the computers would be put under: usually high temperatures in dusty environs, atop a fast moving and turbulent ride.*

And in Polish such examples may be enumerated: *W teorii bytu stoicy byli materialistami, ale dość specyficznego rodzaju. Ich podstawową zasadą w tej dziedzinie było przyjmowanie, że wszystko co istnieje jest materią. Materia ta, inaczej niż w poglądach Epikura, ma charakter ciągły i składa się z mieszaniny bytów. O to co jest celem ostatecznym świata, sprzeczali się poszczególni stoicy, zgadzali się jednak zawsze z tym, że celowi temu nie sposób się przeciwstawić, bo realizuje go wszechobecna Pneuma. Stąd świat ma charakter dość ściśle deterministyczny*.

Or the following example, which illustrates a specific stylistic inconsistency, informality and inadequacy of register: *W kraju, gdzie spośród wszystkich możliwych cech siatkarza najbardziej ceni się warunki fizyczne, takie jak wzrost, czy masa, narodził się gracz, którego warunki fizyczne, choć niezłe, nie powalają na kolana, a który potrafi wykonać każdy element nie gorzej niż jego wyspecjalizowani, znacznie wyżsi i silniejsi fizycznie koledzy z kadry narodowej. Siergiej Tietiuchin, bo o nim mowa, mierzy sobie 197 cm i choć jak na reprezentanta Rosji nie jest to wzrost zbyt wysoki, Tietiuchin jest absolutną gwiazdą i najlepszym zawodnikiem tego zespołu.*

It seems that the authors, familiar with the linguistic norms applying to an encyclopedia entry, strive to achieve a proper level of formality. However, in many cases, they do not maintain stylistic homogeneity, which is visible in the colloquial elements interwoven in the articles.

5. Phatic communion

It is a frequent technique to establish contact with the readers employed by the authors of Wikipedia. This is achieved in both English and Polish editions either by the use of the pronouns *you* or *we* and their Polish equivalents, respectively. In using these forms, the authors wish to involve the reader in the process of explication and to facilitate the acceptance of the specified line of reasoning.

In English, the use of the pronoun *you* amounted to 3%. For instance: *tests that challenge you to restart all the oxygen generators (called Oxyds) on your home planet*; *the marketing method of the Oxyd series was quite unique: you could obtain the complete games for free from shareware-CDs, for example. Then, you could play through the first ten levels without any restrictions*; *when you remove the cap, you can clearly hear gas.*

However, while the use of the pronoun *you* is frequently applied by the English authors, in Polish entries only one instance of such expression was found: *Jeśli lewą dłoń ustawisz tak, aby pole magnetyczne...*

The use of the pronoun *we*, *pluralis modestiae*, is also a frequent means of achieving impersonality, helping to avoid a direct presentation of subjective opinions (it appeared in 4% of the entries). *We* occurs as a bonding of author and reader in a joint enterprise, which, in this case, is the transmittal of information. This form is frequently applied in scientific presentations, explanations and arguments. Due to the use of this technique, entries in Wikipedia in many cases resemble scholarly disquisitions. For instance: *thus, in contrast to the grandeur of his composition at Il Gesu, we see Gaulli gradually adopting less intense colours, and more delicate compositions after 1685 – all hallmarks of the Rococo*; *we are given a collection of subsets S of a universe T and asked to find a subset H of T that intersects ("hits") every set in S. We additionally require that H have at most a given size K.*

Establishing phatic communion in this way is even more common in Polish entries, as such terms were present in 11.5% of the articles. For example: *Wilderness jest największym miejscem, w którym możemy walczyć*; *przypisując każdemu punktowi wartość, mamy funkcję*; *znamy za to cenę, za którą otrzymał wolność*; *nie wiemy jak należy poprawnie odczytać ten znak*; *jeśli chcielibyśmy odczytać to imię poprawnie.*

By the use of these pronouns, authors express their awareness of audience, initiate some form of interaction, as well as indicate shared or common ground with the readers.

6. Repetitions

Specific linguistic ineptitude is visible in frequent repetitions (in English 4.3%, in Polish 3.2%), which could be easily avoided by the application of a synonymous phrase, or by restructuring the sentence. It is clear that authors often encounter problems with finding substitutable means of expression that would help the sentence acquire a more formal style, e.g.: *some similarity to the Times Square New Year's Eve ball drop is presumably intended. The peach tie in is presumably because of Atlanta's affinity for peaches*; *from the above table, clearly there is a problem with the above theory*; *there are clues on many landscapes: some are helpful, but others are confusing or not so helpful*; *he was reportedly arrested and he reportedly murdered one of his inmates*; *it is considered a problem considering that* (...). In Polish: *człowiek to jedna sprawa, a środowisko to sprawa odmienna*; *powoduje to, że wiązanie to*; *dlatego zazwyczaj reagują same ze sobą, a ich stężenie jest zazwyczaj niskie.*

7. Nondescriptiveness

Nondescriptiveness applies in this context to a frequent non-definiteness of description. Authors very often refrain from providing elaborations and specifications of particular information.

In English entries, this lack of precision and exactitude of description is visible in a frequent use of the determiner *some*, which was distinguished in 5% of the articles, for instance: *edges of leaves of some vascular plants*; *announced some decrees that caused public friction between the two*; *some EX city bird planes.*

Similarly, in many cases authors do not quote the sources of the information they provide, are unclear and not specific as far as the data they refer to are concerned, e.g.: *according to some reports*; *some child specialists say*; *some argue*; *there is some evidence.*

Such nondescriptiveness is visible in vagueness of description, lack of application of specific attributes and inability to choose a more determinate means of evaluation. Authors resort especially to the adverb *somewhat*, which also carries a degree of personal judgement, e.g.: *somewhat ironic*; *somewhat older*; *somewhat more complex case.*

In Polish such nondescriptiveness is even more frequent (7%) and is visible in the use of a variety of terms, e.g.: *jakiś elektron nie ma pary*; *jakimiś korzystnymi właściwościami*; *niektórzy uważają*; *jedni twierdzą*; *według innych źródeł*, or: *występują dość powszechnie*; *niejako od końca*; *nieco różne*; *nieco silniejsza.*

The aim of this article is to present the most frequently occurring grammatical and stylistic mistakes of the entries to English and Polish Wikipedias. As stated above, in both encyclopedias, an array of grammatical and stylistic mistakes can be distinguished. It may be stated that informality of expression, abundance of colloquial terms and presence of grammatical mistakes constitute the main stylistic faults of Wikipedia.

The frequency of these faults, as the data prove, is higher in the Polish edition. While colloquial expressions may be distinguished in both editions, the percentage of lexical errors, evaluative expressions and grammatical mistakes is higher in the Polish issue of the encyclopedia. This may be attributed to the fact that the number of moderators and contributors devoted to this edition of the encyclopedia is smaller and, thus, the process of editing and error correction may last longer.

The above mentioned faults occur either due to authors' negligence, lack of editorial control or because of lack of linguistic awareness guiding the choice of a proper register and adequate means of expression useful in a given context. What is more, the outcomes of this study prove that the use of informal register is permeating various contexts of usage previously reserved for formal expressions. However, it should be also stated that within the analyzed material, many of the entries were written in formal register, comparable to the style of other encyclopedias. Many of the most faithful editors of Wikipedia strive to achieve the best results as far as the factual, linguistic and graphic sides of the encyclopedia are concerned. Nevertheless, with the openness of the encyclopedia and with the number of contributors rapidly growing, it seems impossible to avoid the presence of grammatically and stylistically improper entries.

References

Crystal D., Davy D. (1969): *Investigating English Style*, London.

Gajda S. (1990): *Współczesna polszczyzna naukowa. Język czy żargon?* Opole.

Herring S.C., Emigh W. (2005): *Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias*, "Proceedings of HICSS-38", s. 1–10.

Kamińska-Szmaj I. (1990): *Różnice leksykalne między stylami funkcjonalnymi polszczyzny pisanej,* Wrocław.

Lih A. (2004): *Wikipedia as participatory journalism: Reliable sources?* www.jmsc.hku.hk/faculty/alih/publications/utaustin-2004-wikipedia-rc2.pdf

Quirk. R. (1972): *A Contemporary Grammar of English*, London.

Swales J.M. (1990): *Genre Analysis*, Cambridge.

Tereszkiewicz A. (2006): *Analiza gatunkowa encyklopedii internetowej Wikipedia*, „Biuletyn PTJ" LXII, s. 81–91.

Viegas F.B., Wattenberg M., Dave K. (2004): *Studying cooperation and conflict between authors with history flow visualizations*, "Computer Human Interaction 2004", s. 575–582.

Wilkoń A. (1987): *Typologia odmian językowych współczesnej polszczyzny*, Katowice.

Streszczenie

Błędy językowe i stylistyczne w artykułach angielskiej i polskiej encyklopedii internetowej Wikipedia

Artykuł ukazuje wyniki analizy językowej wybranych tekstów angielskiej oraz polskiej encyklopedii internetowej Wikipedia. Forma Wikipedii jest zdeterminowana przez kształt i sposób funkcjonowania internetu. Encyklopedia ta jest tworzona w całości przez ochotników, co wpływa na warstwę językową poszczególnych artykułów. Analiza ma na celu wyłonienie najczęściej spotykanych błędów językowych pojawiających się w encyklopedii. Określa najczęściej pojawiające się błędy leksykalne, gramatyczne oraz stylistyczne, wskazując na frekwencję ich występowania w obu wersjach encyklopedii. W artykule przedstawiono również cechy różnicujące polskie wydanie Wikipedii od wydania angielskiego.