Експериментальна оцінка лінгвістичних аспектів якості технічної документації

Анастасія Колесник

магістр кафедри інтелектуальних та комп'ютерних систем, Національний технічний університет «Харківський політехнічний інститут», Україна, E-mail: kolesniknastya20@gmail.com

Ніна Хайрова

д. т. н., професор, професор кафедри інтелектуальних та комп'ютерних систем, Національний технічний університет «Харківський політехнічний інститут», Україна, E-mail: nina_khajrova@yahoo.com

The object of the research is information and linguistic means of improvement of technical documentation quality and creation of corpus. The purpose of research is the analysis of the existing valuation methods of technical documentation and software development which increases quality of technical documentation by information and linguistic means and software development for count of necessary volume of selection on the basis of the corpus which contains instructions for the user from different vendors.

Ключові слова — комп'ютерні науки, лінгвістичні аспекти, корпус, критерії, генеральна сукупність, вибірка.

I. Вступ

Для створення будь-якого складного технічного продукту необхідна детальна та якісна розробка технічної документації, що пов'язана з цим продуктом [7]. Правильно складена документація є основою функціональності і ефективності інформаційних систем. Щоб максимально допомогти користувачеві, технічний документ має бути граматично і пунктуаційно правильно оформленим. Не менш важливим показником якісності технічного документу є його стиль. Він має відповідати тематиці документу і допомагати виконувати його основні функції. Крім того, кожний технічний документ має відповідати стандартам технічної документації, критеріям якості, що висунуті конкретно до цього документу, та містити усю необхідну інформацію, щоб відповідати вимогам користувача [8]. У той самий час це є і проблемою, оскільки немає окремо виділених критеріїв, спираючись на які, можна впевнено сказати, що документ є якісним. Також немає комп'ютерних програм, котрі можна використовувати для оцінки якості технічного документу. Саме тому якість технічного документу та її оцінка є актуальним питанням у наш час [9].

II.Постановка задачі дослідження

Якісно складений технічний документ має відповідати не тільки базовим стандартам, а і задовольняти вимоги до якості від користувача та власника. Такий документ має бути орієнтованим на конкретне завдання, точним, повним, релевантним, здатним до аналізованості та модифікованості [4]. Для оцінки якості будь-якого документу необхідно виділити критерії цієї оцінки, згідно з якими можна буде стверджувати, що отриманий для аналізу документ є якісним. Оскільки у даній роботі, у якості досліджуваного матеріалу, було обрано тексти інструкцій для користувача, то основні критерії, що необхідно виділити, спираючись на Microsoft Manual of Style, це граматичні, пунктуаційні та стильові [3].

У процесі дослідження були виділені лінгвістично-граматичні, лінгвістично-пунктуаційні та стильові критерії оцінки якості технічної документації і запропоновано алгоритм, що реалізує метод підвищення якості, що використовує виявленні критерії. Алгоритм реалізовано та протестовано на корпусі технічної документації текстів інструкцій, оптимальний розмір якого було вирахувано за допомогою методів статистики виведення корпусної лінгвістики.

У результаті дослідження виявлено 13 лінгвістичних критеріїв, які можуть суттєво вплинути на якість документу. Дані були розподілені на три групи: лінгвістично-пунктуаційні, лінгвістично-граматичні та стильові.

1. Написання чисел від 1 до 9 словами.
2. Написання чисел від 10 і більше цифрами.
3. Написання чисел (навіть тих, що менше 10), які уточнюються виміром, цифрами.
4. Використання від і до (*from* і *through*) замість *between* і *and* щоб описати діапазони чисел.
5. Використання формату MMMM DD, YYYY для запису дати.
6. Відсутність скорочень назв місяців. Замість цього застосування повної форми слова.
7. Застосування курсивного форматування замість написання слова усіма великими літерами.
8. Використання однослівних дієслів замість дієслівних виразів.
9. Використання міжнародних варіантів запису термінів.
10. Використання тільки одного пробілу після розділового знаку.
11. Відсутність коми у форматі дати MMMM YYYY.
12. Застосування розділового знаку (кома, крапка і т. д.) одразу після слова, тобто без пробілу.
13. Відсутність косої риски у конструкціях, що вказують на вибір, такий як *he/she,* у якості замінника *or*.

III.Опис та обґрунтування проведеного експерименту

Для перевірки правильності обраних критеріїв, було розроблено спеціальний корпус, що повинен вилучати вибіркові дані, які містять феномени, що підлягають лінгвістичному опису.

У репрезентативний корпус було введено тексти інструкцій до 3 видів електроприладів: холодильник, мікрохвильова піч і посудомийна машина. Було обрано 4 фірми виробники, такі як: Bosch, Beko, Indesit, Electrolux. Вибір цих інструкцій зумовлений тим, що нині вони є лідерами з виробництва таких пристроїв як мікрохвильові печі, холодильники і посудомийні машини. Таким чином наповнення корпусу стає актуальнішим.

Оскільки розробити реальний повний корпус (генеральну сукупність) на цю тематику не є можливим, то для подальшого дослідження потрібно використати та обґрунтувати репрезентативну вибірку. Для цього застосовується вибірковий підхід. З генеральної сукупності, властивості якої підлягають дослідженню, певними методами формують вибірку (обмежена кількість об'єктів), до яких застосовують дослідницькі методи. У результаті методів спостережень, експериментальних дій і вимірів над об'єктами вибірки отримують емпіричні дані. Обробка емпіричних даних за допомогою методів описової статистики дає показники вибірки. Застосовуючи методи статистичного виводу до статистик, отримують параметри, що характеризують властивості генеральної сукупності.

Вибіркова сукупність (вибірка) − це відібране за суворо заданим правилом певне число елементів генеральної сукупності. Вибірка є своєрідною мікромоделлю усієї генеральної сукупності, тобто за усіма основними якісними характеристиками, що вивчаються, і контрольними ознаками вона буде своєю структурою максимально повторювати структуру генеральної сукупності [5].

У такій репрезентативній вибірці є усі елементи генеральної сукупності, наприклад, об'єкти, що часто проявляються в генеральній сукупності, частіше проявляються і в ній. Основний принцип побудови вибірки полягає в тому, щоб усі елементи генеральної сукупності мали рівні шанси потрапити у вибірку. Але як би ретельно не дотримувалися цього принципу, випадкові помилки все ж матимуть місце у вибірці [6].

Друга проблема вибіркового дослідження полягає в проблемі оцінки, пов'язаної з тим, що висновки, що робляться на основі даних вибірки, адекватно характеризують тільки властивості вибірки, а їх перенесення на властивості генеральної сукупності призводитиме до деякої погрішності. Проблема оцінки полягає в необхідності використання з максимально можливою надійністю результатів, отриманих по вибірці для висновків про генеральну сукупність.

Базуючись на дослідженнях, проведених у роботах [1,2], для визначення оптимального об'єму вибірки будемо використовувати долю ознаки (долю речень з помилками, які не відповідають критеріям оцінки якості технічної документації) в генеральній сукупності за відповідною вибірковою характеристикою.

Нехай доля ознаки в генеральній сукупності, яка показує відношення числа релевантних документів до загального числа документів в генеральній сукупності, рівна R. Вибіркова оцінка Rs долі R рівна Rs = M/N, де, N - об'єм досліджуваної вибірки, а M - кількість виявлених в ній релевантних документів. Оскільки об'єкти випадково відбираються у вибірку, то вибіркова доля Rs може набувати будь-яких значень в інтервалі [0;1], причому Rs=0, коли жоден релевантний документ не потрапив у вибірку і Rs=1, якщо усі документи у вибірці релевантні [2].

Оскільки Rs, а, отже, і помилка вибірки є випадковими величинами з одним і тим же розподілом вірогідності, введемо породжувану конкретною точковою оцінкою Rs інтервальну оцінку, у межах якої з деякою довірчою імовірністю P лежатиме доля ознаки генеральної сукупності [2].

Довірча імовірність P показуватиме імовірність того, що інтервальна оцінка містить в собі невідому долю ознаки генеральної сукупності. Оскільки в загальному випадку розподіл величини R несиметричний, то інтервальна оцінка, або довірчий інтервал, випадкової величини R має вигляд:

$$P(R_S - E_1 < R < R_S + E_2) = 1 - \alpha, \quad (1)$$

де $[R_S - E_1; R_S + E_2]$ –довірчий інтервал, $R_S - E_1$; $R_S + E_2$ - довірчі межі, $P = 1 - \alpha$ – довірча вірогідність, α – рівень значущості.

У разі симетричного розподілу R довірчий інтервал також симетричний відносно величини R і має вигляд: P (|R - Rs|< E) = 1 - α , де величина E - гранична помилка, що характеризує точність вибірки.

Довірчий інтервал для долі ознаки потрібно визначати, строго кажучи, базуючись на біноміальному законі розподілу. Починаючи з вибірок об'ємом не менше 20, біномний розподіл добре апроксимується нормальним розподілом з параметрами: середнє <Rs> = R, дисперсія D (Rs) = R (1 - R) /N, стандартне відхилення $\sigma(R_S) = [D (RS)]1/2]$. При цьому довірчий інтервал може бути розрахований по формулі $P(|R - R_S| < E_\alpha) = 2\Phi(Z_\alpha) = 1 - \alpha$, де $\Phi(Z_\alpha)$- функція Лапласа. Гранична помилка вибірки знаходиться при цьому з рівності $E_\alpha = =Z_\alpha\sigma(R_S)$ [2].

$$P(|R - R_S| < E_\alpha) = 2\Phi(Z_\alpha) = 1 - \alpha, \quad (2)$$

$$P(|R - R_S| < E) = 1 - \alpha, \quad (3)$$

$$E_\alpha = =Z_\alpha\sigma(R_S) \quad (4)$$

$$\sigma(R_S) = [D(R_S)]^{1/2} \quad (5)$$

$$E_\alpha = Z_\alpha [R(1 - R)/N]^{1/2} \quad (6)$$

$$D(R_S) = R(1 - R)/N \quad (7)$$

$$|R - R_S| < Z_\alpha [R(1 - R)/N]^{1/2} \quad (8)$$

За величину довірчої вірогідності зазвичай вибирають значення 0,95 (тоді рівень значущості α = 0,05). У такому випадку $2\Phi(Z_\alpha)=0,475$ — функція Лапласа. При цьому $Z_{0,05} = 1,96$.

Для того, щоб визначити об'єм потрібної нам вибірки при заданій довірчій імовірності і граничній помилці, замінимо |R - RS| на E і розрахуємо рівняння, відносно N.

$$N = [Z^2 R_S(1 – R_S)]/E^2 \quad (9)$$

У співвідношення входить вибіркова доля RS для визначуваного ще невідомого об'єму вибірки. Оскільки ця доля невідома, розумно визначити її так, щоб об'єм вибірки N був максимальним (тобто годився при усіх допустимих RS). Не важко вирахувати , що максимум N як функції Rs досягається при Rs = ½, тобто $Nmax = Z^2/4E^2$.

У нашому випадку у результаті роботи програми, рахується доля ознаки як по усьому корпусу так і по кожному текстовому файлу окремо. Тож потрібно використовувати отриману величину у формулі 9. При цьому необхідний об’єм вибірки буде менший за максимальний, що є правильно.

Досить часто при знаходженні доль ознак використовується величина граничної помилки E = 0,05. Якщо використовувати E = 0,05 при розрахунку $Nmax = Z^2/4E^2$, то максимальний обсяг вибірки при R_S = ½ буде дорівнювати 384,16. Але в рамках нашого дослідження також розглядалися результати підрахунку необхідного обсягу вибірки при граничній помилці від 0,1 до 0,01.

Висновок

Проведений аналіз дозволив визначити лінгвістично-граматичні, лінгвістично-пунктуаційні та стильові критерії оцінки якості технічної документації.

Інформаційно-лінгвістичні критерії якості відповідають за правильну подачу і оформлення інформації — згідно з правилами пунктуації та граматики. Ці критерії тісно пов’язані зі стилем та завданням (для чого створений) документа. Лінгвістично-пунктуаційні критерії, головним чином, зосереджуються на проблемах пунктуації, які є питанням зв'язку стилю технічної документації з таким лінгвістичним аспектом. За граматичну правильність тексту технічного документу відповідають правила граматики, з яких було виокремлено критерії, що можуть суттєво вплинути на якість технічної документації [3].

Особливим аспектом оцінки технічного документа є його стильове оформлення. За допомогою послідовного, чітко визначеного стилю, з правильно оформленими числами, датами й одиницями виміру, зміст технічного документа стає зручнішим для прочитання і легшим для розуміння. Тож визначення якості стилю технічного документу є одним із першочергових завдань під час його оцінки.

Також розроблено алгоритм, що реалізує метод підвищення якості, у якому використовуються виявлені критерії. Його було протестовано на корпусі, що складається з текстів інструкцій, та вирахувано оптимальній розмір корпусу — 20 000 тисяч слів. На його основі вирахувано оптимальний розмір вибірки, а також здійснено тестування виявлених у процесі дослідження критеріїв. Згідно з ними, у текстах корпусу знаходилися і виправлялися помилки.

Для того щоб висновки за вибіркою можна було віднести і до генеральної сукупності, рахується необхідний обсяг вибірки, тобто її оптимальний розмір для підрахування якогось критерію (долі ознаки). У межах цього дослідження за долю ознаки було обрано критерії якості технічного документу, тому вона показує відношення речень, що містять помилки (граматичні, пунктуаційні та стильові) до загальної кількості речень у корпусі.

За результатами підрахувань доля ознаки, тобто Rs, за корпусом дорівнює 0.144. Завдяки цій величині та програмному продукту було підраховано необхідний розмір вибірки, що при E=0.05, буде дорівнювати 189.84. При такому ж E та при Rs = ½, N_{max}=384.16. Отже, можна зробити висновок, що результати підрахунків (Рис. 1) є достовірними, оскільки необхідний обсяг вибірки має бути меншим за максимальний.

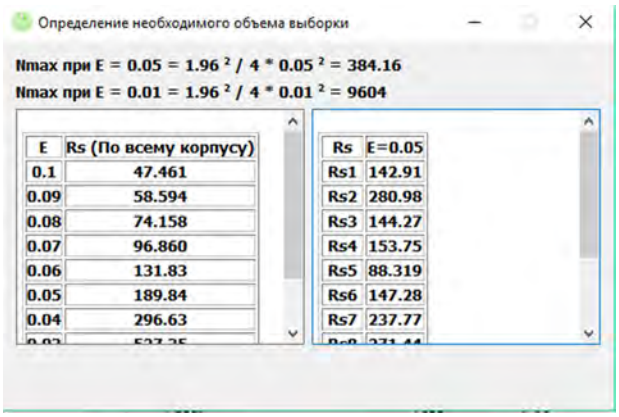

Рис. 1 Результати підрахунків

Відповідно до отриманих результатів, можна стверджувати, що за допомогою такого продукту можна суттєво покращити якість технічних документів та розширити межі оцінки.

Література

1. Khairova N. Evaluating Effectiveness of Linguistic Technologies of Knowledge Identification in Text Collections / Khairova N., Shepelyov G., Petrasova S. // Transactions on Business and Engineering Intelligent Applications. ITHEA. – Rzeszow – Sofia. 2014. – P. 71-75

2. Хайрова Н. Решение проблемы формальной оценки эффективности технологий идентификации знаний в слабоструктурированной текстовой информации / Н.Ф. Хайрова, Н.В. Шаронова, Д.Ю. Узлов // International Journal “Information Content & Processing” – 2014. – С. 9.

3. Microsoft Manual of Style 4th edition / Published by Microsoft Press. – 2012. – 439 p.
4. Guidelines for writing documentation: [Електронний ресурс]. – Режим доступу: http://bluebream.zope.org.
5. Выборка и доверительные интервалы [Електронний ресурс].– Режим доступу: http://www.mathelp.spb.ru/book2/tv12.htm.
6. Основные задачи и методы математической статистики: [Електронний ресурс]. – Режимдоступу:http://uchebnikionline.com/statitika/matematichna_statistika_htm.
7. Wingkvist A. A Visualization-based Approach to Present and Assess Technical Documentation Quality / Anna Wingkvist, Morgan Ericsson. – Sweden : Linnaeus University. – 2011. – 10 p.
8. Juran J. Juran's Quality Control Handbook / 5th ed., McGraw-Hill. – 1998. – 1136 p.
9. Blicq R. Guidelines for Writing English Language Technical Documentation for an International Audience / Ron Blicq. – INTECOM International Language Project Group. – 2003. – 40 p.

Функціональні особливості вживання модальних дієслів в англомовних юридичних текстах: параметри частоти та сполучуваності

Ірина Карамишева

к. філол. н., доцент, доцент кафедри прикладної лінгвістики, Національний університет «Львівська політехніка», Україна, E-mail: iryn_ka@ukr.net

The presented research focuses on the functional peculiarities of modal verbs as means of modality expression in legal discourse with the aim to define the frequency of their usage in distribution agreements. The material selected for the research consists of 100 English contracts as corpora for the analysis. All the analyzed documents have been taken from the Internet sourses. The average length of each of the analyzed texts is about 17,000 words in one text. On the basis of the material presented in the theoretical English grammars, basic modal verbs were selected and became the object of study. The examples of modal verbs usage in the distribution agreements were analyzed, which allowed describing their basic functional characteristics as well as combinability patterns. It was found that the value of modal verbs is created by context and grammatical structures in which they are used. The frequency of occurrence of specific modal verbs in legal texts is different from their usage in common everyday language. The analysis leads to the conclusion that there is a need to conduct further studies that will determine exact translations of each modal verb and the morphological and syntactic forms they take.

Ключові слова — модальність, модальні дієслова, функціональні характеристики, частота вживання, юридичний дискурс, договір дистрибуції.

I. Вступ

У сучасному комерціалізованому суспільстві зростає обсяг документообігу, в тому числі існує постійна потреба написання контрактів (договорів), постає необхідність автоматизованого/автоматичного опрацювання мовних даних та створення шаблонів укладання офіційної документації як прикладного лінгвістичного завдання. Сучасною тенденцією лінгвістичних досліджень, спрямованих на вивчення офіційної документації, є дослідження особливостей функціонування певних мовних одиниць, які є доволі частотними, а також структур, у яких вони можуть вживатись. Такі структури можна піддавати формалізованому аналізу та використовувати у прикладних лінгвістичних програмах, що і спричиняє **актуальність** пропонованого дослідження.

Офіційна документація, в тому числі контракти, договори, характеризуються специфікою мовного представлення. Специфічним є не лише підбір лексики, що характеризує мову права, а й втілення певних категорій граматичними засобами, зокрема йдеться про вираження модальності в юридичному дискурсі. Відтак **об'єктом** пропонованого **дослідження** є модальні дієслова як засоби вираження модальності в англомовному комерційному контракті (договорі дистрибуції). **Предмет дослідження**: функціональні особливості модальних дієслів в англомовному юридичному дискурсі (на матеріалі договорів дистрибуції). Загальною **метою дослідження** є вивчення функціональних характеристик модальних дієслів в юридичному дискурсі та визначення частоти їхнього вживання та особливостей сполучуваності.

II. Формування фактологічної бази дослідження

Матеріал дослідження включає: набір комерційних контрактів англійською мовою, а саме договорів дистрибуції (100 примірників), оскільки вони є одними з найпоширеніших видів комерційних контрактів. Електронний ресурс: contracts.onecle.com/type/147.shtml

Утворену вибірку модальних дієслів з англомовних договорів дистрибуції проаналізовано на предмет частоти їхнього вживання в такому виді англомовних юридичних текстів, а також визначено особливості вживання з