УДК 004.912

О. А. Лозицький, Н. Е. Кунанець
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

СИСТЕМА ОПРАЦЮВАННЯ ТЕХНІЧНИХ ТЕКСТІВ УКРАЇНСЬКОЮ МОВОЮ З МЕТОЮ ЇХ АДАПТАЦІЇ ДЛЯ ЛЮДЕЙ З ВАДАМИ ЗОРУ

Розглянуто методи і засоби розроблення системи опрацювання технічних текстів українською мовою для забезпечення потреб інформаційної доступності незрячих людей.

Ключові слова: вада зору, особливі потреби, програмна система, обробка україномовного технічного тексту, математична формула, синтаксичне дерево, озвучення, автоматизоване робоче місце, DAISY, MathML.

This paper reviews methods and means of development of the applied programmed system of Ukrainian technical texts processing to meet the needs of information accessibility for the blind people.

Keywords: visual impairments, blind, special needs, applied programmed system, Ukrainian technical text processing, mathematical formula, visual impairment, syntax tree, sounding, automation equipped working place, DAISY, MathML.

Вступ

У сучасному суспільстві інформація як один із основних ресурсів інтегрується в соціокомунікаційне середовище, стає основною його складовою, необхідним атрибутом забезпечення діяльності усіх напрямів його функціонування. Її якість і достовірність, оперативність отримання покладено в основу численних рішень, що приймаються на різних рівнях управління. В процесах поширення інформації особливу роль відіграють бібліотеки, на які покладається функція інформаційного забезпечення різних категорій користувачів.

Бібліотека є одним із найдавніших соціальних інститутів суспільства, що містить інформаційні та культурні компоненти й забезпечує стійкість комунікаційних зв’язків і відносин. Як кожен соціальний інститут, вона розвивається, і цей розвиток пов’язаний зі змінами суспільних потреб. Відповідаючи за виробництво, зберігання і трансляцію соціально значущої інформації, вона пройшла складний шлях становлення, розвитку та адаптації до конкретних суспільно-історичних умов. Сьогодні книгозбірня перетворюється на активного учасника комунікативного процесу, що ґрунтується на вдосконалених не лише бібліотечних, а й соціально-комунікативних технологіях, що набуло в сучасних умовах ключового сенсу, зокрема в контексті надання інформації для осіб з особливими потребами. Для того, щоб ефективно обслуговувати користувачів з обмеженими можливостями, потрібно переосмислити традиційні форми бібліотечно-інформаційної роботи. Особливо це стосується форматів представлення інформації для осіб з особливими потребами в зручній для її сприйняття формі.

У цьому контексті неминучою вимогою сьогодення є розроблення принципово нових технологій доступу до інформаційних ресурсів, програмного забезпечення, орієнтованого саме на таку категорію користувачів. Для цього потрібно на основі системного підходу розробити технології обслуговування користувачів з особливими потребами, що ґрунтуються на новітніх методологіях інформаційно-бібліотечного обслуговування з використанням сучасних спеціальних програмно-алгоритмічних та комп’ютерно-технологічних комплексів.

Аналіз останніх досліджень та публікацій

У працях дослідників із розвинених країн, зокрема США та держав Європи розглядаються питання автоматизованого подання текстів, адаптованих для сприйняття незрячими. Р. Е. Ладнер, М. І. Івори, Р. Рао, С. Бургсталер, Д. Комден, С. Ган, М. Ренцелманн, С. Кріснанді, М. Рамасами, Б. Слабоський, А. Мартін, А. Лаценські, С. Олзен, Д. Кроце [1] розробили технологію подання шрифтів Брайля як графічних об'єктів. Аналогічні дослідження проводять інші науковці, зокрема Ц. Жаянт, М. Ренцелманн, Д. Вен, С. Кріснанді, (S. Krisnandi), Р. Ладнер, Д. Комден [2] та Н. Амікк, Я. Коркоран [3]. Поряд з цим розроблено якісні системи синтезу мови [4], програми читання тексту з екрана [5], програмне забезпечення для озвучення технічних текстів [6], спеціально автоматизовані робочі місця, навчально-методичні матеріали тощо. Ці технології практично неможливо автоматично адаптувати до текстів українською мовою через специфіку звучання, своєрідну фонетику та граматику, правила побудови речень тощо.

Мета статті – проаналізувати можливості розробленого програмного забезпечення, що дає змогу створювати мультимедійний інформаційний контент для осіб з вадами зору з метою інформаційного забезпечення навчального процесу учнів та студентів з проблемою зору.

Особливої актуальності реорганізаційні зміни в роботі бібліотек набувають в контексті інформаційного супроводу процесу навчання осіб з особливими потребами. Незважаючи на наявність значної кількості педагогічних, програмних і технічних засобів для полегшення процесу навчання людей з вадами зору, інформаційне суспільство висуває вимоги подальшого розвитку в цьому напрямі. Процес інформаційного супроводу неможливий без комп'ютерного подання та адаптації навчально-методичного матеріалу до потреб цієї категорії користувачів.

Автоматизації процесу створення технічних та природничих книг українською мовою

Засадничим є принцип рівних можливостей, тобто надання того самого обсягу та якості інформаційних послуг, якими користуються усі інші громадяни, на засадах інклюзії.

У цьому контексті виникає потреба формування електронної бібліотеки, інформаційні ресурси якої адаптовані для потреб таких членів суспільства.

Поняття електронної бібліотеки (digital library) для осіб з особливими потребами можна окреслити як інтегровану бібліотечно-інформаційну систему, яка дає змогу нагромаджувати, зберігати та забезпечувати доступ користувачів до різних колекцій електронних мультимедійних та повнотекстових документів, поданих у зручному для них форматі із врахуванням комунікаційних каналів сприйняття інформації. Електронна бібліотека забезпечує доступ до віддалених, різнорідних і розподілених інформаційних ресурсів та постійний інформаційний супровід навчального процесу .

Інформаційні ресурси такої бібліотеки повинні обов'язково містити книги у DAISY [7] форматі. Подання інформаційного ресурсу у зазначеному форматі дає змогу його адаптувати для ефективного інформаційного доступу для осіб з вадами зору. Сьогодні можна констатувати гостру потребу в адаптованому для осіб з вадами зору інформаційному контенті українською мовою. Якщо тексти, що подають інформацію лінійно, скажімо, художні твори, які завдяки різноманітним грантовим програмам поступово накопичуються в електронних бібліотеках України, то навчальні матеріали у галузі фізико-математичних та технічних наук, адаптовані для сприйняття незрячими користувачами, практично відсутні. Найбільшу проблему при автоматизованому перетворенні у DAISY формат текстів, що подають відомості з цих галузей наук, наявність формул та математичних виразів. Жоден з наявних програмних продуктів не міг забезпечити правильне їх аудіовідтворення. Для автоматизації процесу створення технічних та природничих книг українською мовою у DAISY форматі розроблено прикладну програмну систему. Ця система забезпечує опрацювання україномовних технічних текстів з метою їх адаптації для людей з вадами зору. Універсальний алгоритм створення книги у форматі DAISY подано на рис. 1. В основу розробленої програмної системи покладено модульну структуру, яка дає змогу реалізовувати її у вигляді відповідних функціональних модулів. Основними компонентами прикладної програмної системи є: драйвери спеціального обладнання, спеціальне та базове програмне забезпечення. Функціональність системи автоматизованого створення книг у DAISY форматі розглянуто у попередніх статтях [8]. У цій статті розглянемо модулі, що входять до її складу і забезпечують аудіоподання математичних виразів та формул. На рис. 2 у блоці (Спеціальне програмне забезпечення) їх виділено штрих-пунктиром.

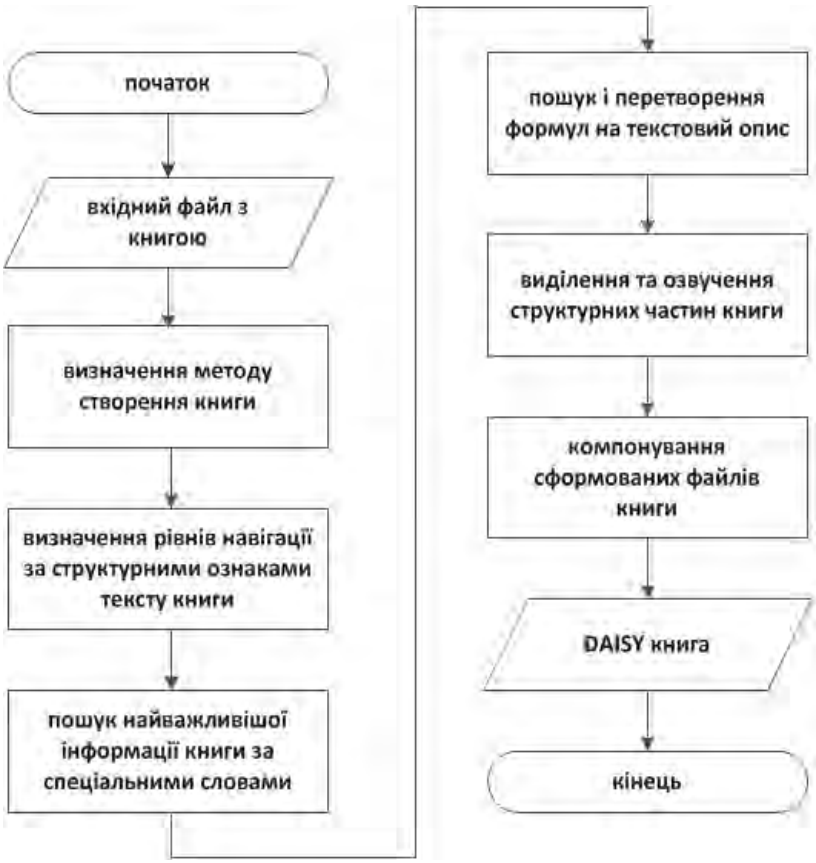


Рис. 1. Блок-схема алгоритму створення DAISY книги

Модуль опрацювання файлів різних форматів.

Система надає змогу користувачеві завантажити файл з електронним документом. Відбуваються аналіз та класифікація вхідних даних із використанням методу класифікації документів. Модуль опрацювання файлів різних форматів надає користувачеві можливість перетворення формату вхідної книги відповідно до вимог системи.

Модуль забезпечує автоматизоване застосування методу перетворення форматів файлів. Сформована у системі база знань подає користувачу підказки щодо застосування основних алгоритмів перетворення форматів вхідних файлів на формат системи HTML/XHTML. Розроблений модуль перевіряє вхідний формат файла і виводить користувачеві повідомлення з подальшими інструкціями щодо конвертування формату вхідного файла. У системі описано варіанти конвертації найпоширеніших форматів електронних документів, зокрема, TXT, DOC, RTF, HTML, PDF, DJVU. У систему завантажується файл текстового або графічного формату, потім система ініціалізує формат вхідного файла, визначає оптимальний метод розроблення книги та виводить повідомлення про необхідність зміни формату вхідного файла.

Наприклад, якщо в систему завантажено файл у форматі текстового редактора MS Word, то користувач отримує повідомлення про те, які кроки необхідно виконати для перетворення вхідного файла на формат HTML/XHTML:

1. Відкрити Ваш документ засобами MS Word або MS Office.
2. Зайти в меню: Файл – Зберегти як.
3. У рядку Тип файла потрібно вибрати Веб-сторінка.
4. Зберегти файл у новому форматі.
5. Завантажити отриманий файл у систему.

Рис. 2. Структурна схема прикладної програмної системи опрацювання україномовних технічних текстів для людей з вадами зору

Модуль пошуку ключових слів у тексті

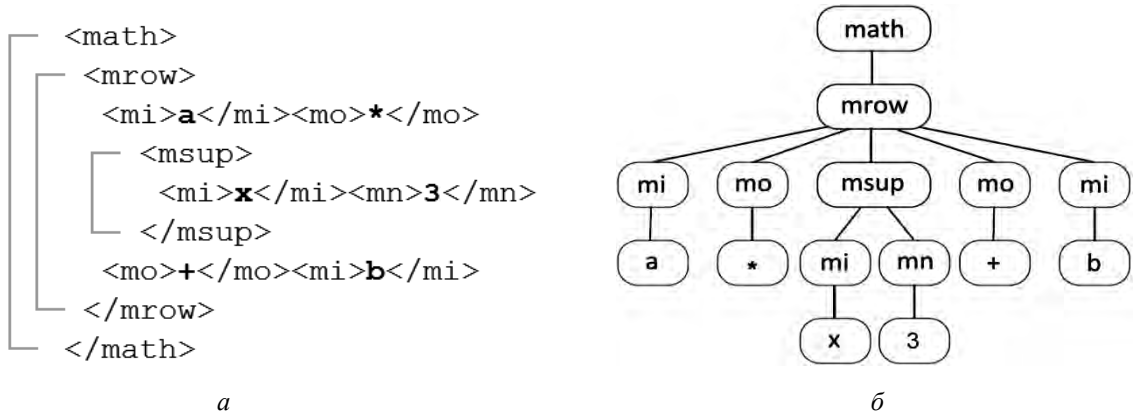
Модуль пошуку ключових слів у тексті забезпечує пошук та маркування тексту за структурними ознаками книги, а також навігацію із визначенням наявності рисунків, найважливішу інформацію книги тощо. Пошук необхідної інформації у тексті книги відбувається із використанням змішаних методів та методу пошуку за ключовими словами. Розроблений модуль знаходить ключові слова в HTML/XHTML файлі за маскою пошуку, відповідно до сформованої бази ключових слів та маркує потрібні фрагменти тексту, що надалі озвучують голосом іншого диктора. Отже, система пропонує користувачеві можливість знайти у книзі необхідну інформацію. Якщо це необхідно, то система шукає ключові слова, задані за допомогою маски пошуку.

Модуль накладання навігації на книгу

Модуль накладання навігації на книгу надає можливість автоматизованого визначення структури документа та накладання навігації із застосуванням методу пошуку за ключовими словами та методу дерев рішень. Відповідно до розробленої класифікації типів книг та побудованого дерева рішень розроблений модуль шукає у вхідному файлі структурні ознаки тексту, визначає, на скільки рівнів необхідно поділити книгу та здійснює поділ. Для ефективного накладання навігації розроблено класифікацію книг та електронних документів за типами, на основі якої цей модуль здійснює накладання навігаційної схеми на вхідний документ і поділяє його на окремі структурні частини. Після завантаження файла система шукає ключові слова відповідно до маски пошуку і на основі результату пошуку визначає, на скільки навігаційних рівнів доцільно поділити книгу. Наприклад, якщо у тексті знайдено ключові слова «Розділ», «Підрозділ», «Зміст» і «Список літератури», система поділить книгу на чотири розділи. В результаті структурні частини тексту маркують для подальшого зберігання в окремі текстові файли та озвучення синтезатором української мови.

Модуль перетворення формул на текстовий опис

Модуль перетворення формул на текстовий опис забезпечує пошук у тексті та вербальне подання математичних формул та спеціальних символів українською мовою відповідно до розроблених правил. Фактично завдання озвучення формули зводиться до розроблення системи правил, за якими внутрішнє подання формули перетворюється на зовнішнє подання українською мовою. Для перетворення MathML на текст використовується модель трансформації синтаксичного дерева [9]. Розроблено спеціальну систему правил для перетворення математичних формул, які подано у різних варіантах запису MathML (презентаційний та семантичний), на текст українською мовою. Така система правил складається із правил запису математичних символів, операторів, а також загальних та уточнених виразів. Правила для уточнених виразів потрібні у тому випадку, коли результат читання залежить не лише від певного вузла дерева, але й від значення його нащадка. Наприклад, x^2 правильно прочитати «ікс квадрат», а не «ікс у степені два». Правила сформовано так, що вихідний текст може бути прочитаний синтезатором української мови.



а *б*

Рис. 3. Запис формули $a = x^3 + b$ *мовою MathML (а) та її зображення у вигляді дерева (б)*

Для трансформування дерево математичного виразу MathML треба подати множиною вузлів $S = \{S_1, S_2, ..., S_n\}$ та функцією для відображення вузлів на список їх нащадків $C : S \to L$, де L – множина всіх можливих списків, сформованих на множині S. Атрибутом $Text(s)$ кожного вузла s є текст, який зберігається у вузлі. Цей текст може бути тегом MathML, числовим записом, назвою математичної змінної або математичним символом.

Для трансформування вузлів застосовують такі правила:

$$r = < t, n, T, G >; \quad (1)$$

$$R = \{r_i\}, \quad (2)$$

де t – значення атрибута $Text(s)$ вузла s, до якого можна застосовувати правило; n – кількість нащадків вузла s; $T = < t_1, t_2, ..., t_n >$ – список довжини n, який задає вимоги до кожного вузла-нащадка; G – шаблон рядка, що генерується правилом; R – множина правил.

Шаблон G може містити посилання на значення вузлів-нащадків, наприклад, «*вираз1* помножити на *вираз2*».

Усі правила трансформування записано у список *RuleList*, а також впорядковано від найбільш деталізованих до найзагальніших. Правило трансформування (1) можна застосувати до вузла s, якщо виконуються такі вимоги:

$$Text(s) = t, \quad (3)$$

$$|C(s)| = n, \quad (4)$$

$$\forall i \in \{1, 2, ..., n\} : t_i = '*' \vee Text(C(s)[i]) = t_i, \quad (5)$$

де $C(s)[i]$ позначає i-й елемент списку $C(s)$.

У результаті застосування правила трансформування до вузла s текстовий атрибут *Text*(*s*) замінюється на рядок, який згенерований за шаблоном *G*, а усі вузли-нащадки вузла *s* видаляються.

Для автоматизації процесу перетворення формул, записаних мовою MathML, на текст розроблено даткові програмні засоби декомпозиції формул і перетворення їх на текстовий опис. Пошук формул у тексті книги відбувається на основі удосконаленого методу трансформації синтаксичного дерева та методу пошуку за ключовими словами. Розроблений модуль шукає математичні формули у документі, відповідно до маски пошуку. Математичні формули в документі HTML/XHTML починаються із тегу $і закінчуються тегом$. У результаті всі теги, які розміщені між цими двома, озвучується. Отже, математична формула, яка записана мовою математичної розмітки MathML, відповідно до методу трансформації синтаксичного дерева і розроблених правил, перетворюється на текстовий опис українською мовою і записуються у текст документа. Цей процес є трудомістким, якщо у документі трапляються формули у графічних форматах. Систему налаштовують так, щоб знайдені математичні формули озвучувались україномовним синтезатором і зберігались в аудіоформаті MP3 або WAV.

Під час розроблення цієї системи враховували, що у електронних документах для запису математичних формул використовується велика кількість електронних форматів та способів їх подання: TeX, MathType Equation, OpenOffice Math, MathML та інші. У деяких документах формули можуть зберігатися у вигляді растрових або векторних зображень (див. таблицю).

Кодування математичних формул у різних нотаціях

Кодування формули (нотації)	Приклад
Традиційне	$1+\sqrt{\frac{x^2-y^2}{x+y}(x-y)}=0$
LaTeX	1+\sqrt{ ((x**2–y**2)/(x+y))*(x–y) }=0
AMS	1+(((x**2–y**2)/(x+y))*(x–y))//2=0
Nemeth	#1+>?X^2"–Y^2"/X+Y#(X–Y)].K #0

Таке розмаїття спричинене незалежним розвитком систем редагування формул, який передував прийняттю єдиного стандарту, до якого різні математичні редактори можуть конвертувати формули. Саме це розмаїття і було певною перешкодою на початкових етапах розроблення прикладної програмної системи. Різноплановість подання формул слід було обов'язково враховувати і розробити модуль, який забезпечує конвертування формул в уніфікований формат. Для цього у системі розроблена база знань з описом покрокових перетворень математичних формул різних форматів (MathType, TeX, LaTeX, JPG, PNG тощо) до мови математичної розмітки MathML.

DAISY MathML вимагає, щоб елементи <math> містили два атрибути:

- altimg – URL адресу відповідного рисунка формули;
- alttext – текстовий опис MathML формули, який можна використати для озвучення.

Використання мови MathML у стандарті DAISY забезпечило підтримку математики у книгах, що «розмовляють», і дало змогу незрячому користувачеві вивчати математику в зручний спосіб.

На рис. 4 зображено, як саме відбувається доступ до відповідної формули під час читання DAISY книги. Елемент <text> файла SMIL, який відповідає запису формули мовою MathML (стрілка 1, рис. 4), використовує атрибути, які вказують на те, що у тексті книги виявлено формулу. Атрибути «altimg» та «alttext» елемента <math> використовують тоді, коли програвач не може прочитати формулу, записану мовою MathML.

Після цих атрибутів у файлі DTBook записується математична формула $\sqrt[3]{x}$ засобами MathML:

```
…
  <m:root>
  < m:mi>x</ m:mi>
  < m:mn>3</ m:mn>
 </ m:mroot>
```

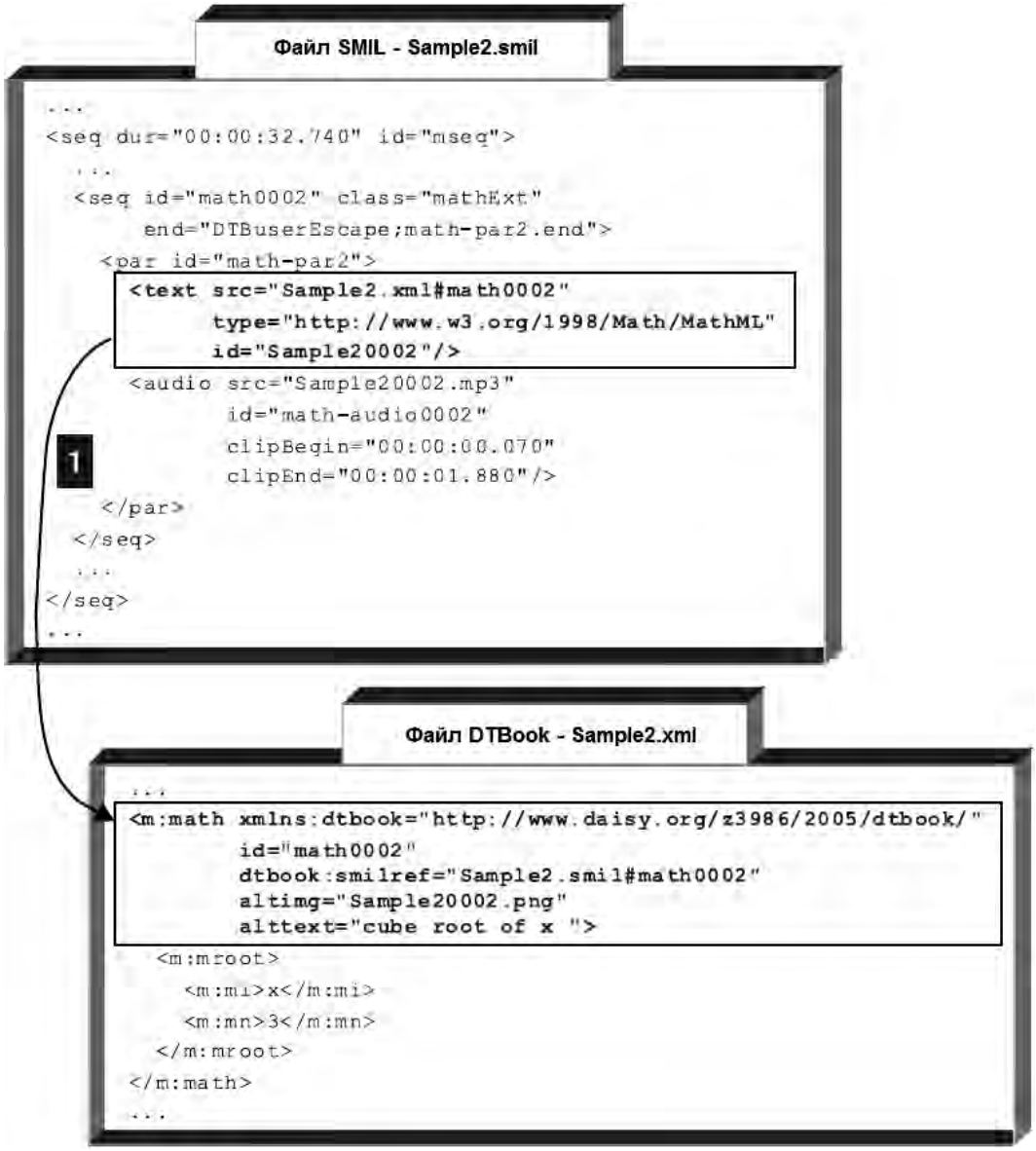


Рис. 4. Використання математичної формули у DAISY книзі

Модуль компонування та зберігання контенту

Модуль компонування та зберігання контенту відповідає за виділення та озвучення основних частин документа, компонування сформованих файлів книги. Цей модуль базується на використанні програмного забезпечення для побудови DAISY книг. Для побудови DAISY книги українською мовою імпортуються створені аудіофайли (MP3 або WAV) у стандартну програму формування DAISY книг. Відбувається запис окремих структурних частин книги, математичних формул, а також найважливішої інформації в окремі текстові файли, озвучувані за допомогою синтезатора української мови «український голос UkrVox – Ігор». Проведений нами аналіз українських голосів дозволяє стверджувати, що «український голос UkrVox – Ігор» є чи не єдиним безкоштовним Speech API голосом середньої якості звучання, який інтегрований нами під час розроблення програмного забезпечення доступності незрячих до аудіоматеріалів. Microsoft Speech Application Programming Interface (Speech API, SAPI) – бібліотека програм для Windows, що забезпечує розпізнавання і синтез голосу у програмах для цієї операційної системи [10]. Потім забезпечується збереження структурних частин тексту книги в окремі файли, а також можливість озвучування їх українською мовою із застосуванням методу синтезу звуку. Розроблений модуль відповідає за зберігання частин вхідного документа у форматах TXT, RTF, DOC, а також під'єднання бібліотеки Windows API для використання синтезатора мови та озвучення тексту.

Якщо необхідно, отримані аудіофайли редагують та нормалізують в аудіоредакторі. Компонування результуючих аудіофайлів і побудова україномовної книги, що «розмовляє», відбуваються у програмному забезпеченні для створення DAISY книг, наприклад, PRS Pro.

Розробленою системою може користуватись зряча людина або незрячий користувач зі спеціальними навичками роботи на комп'ютері.

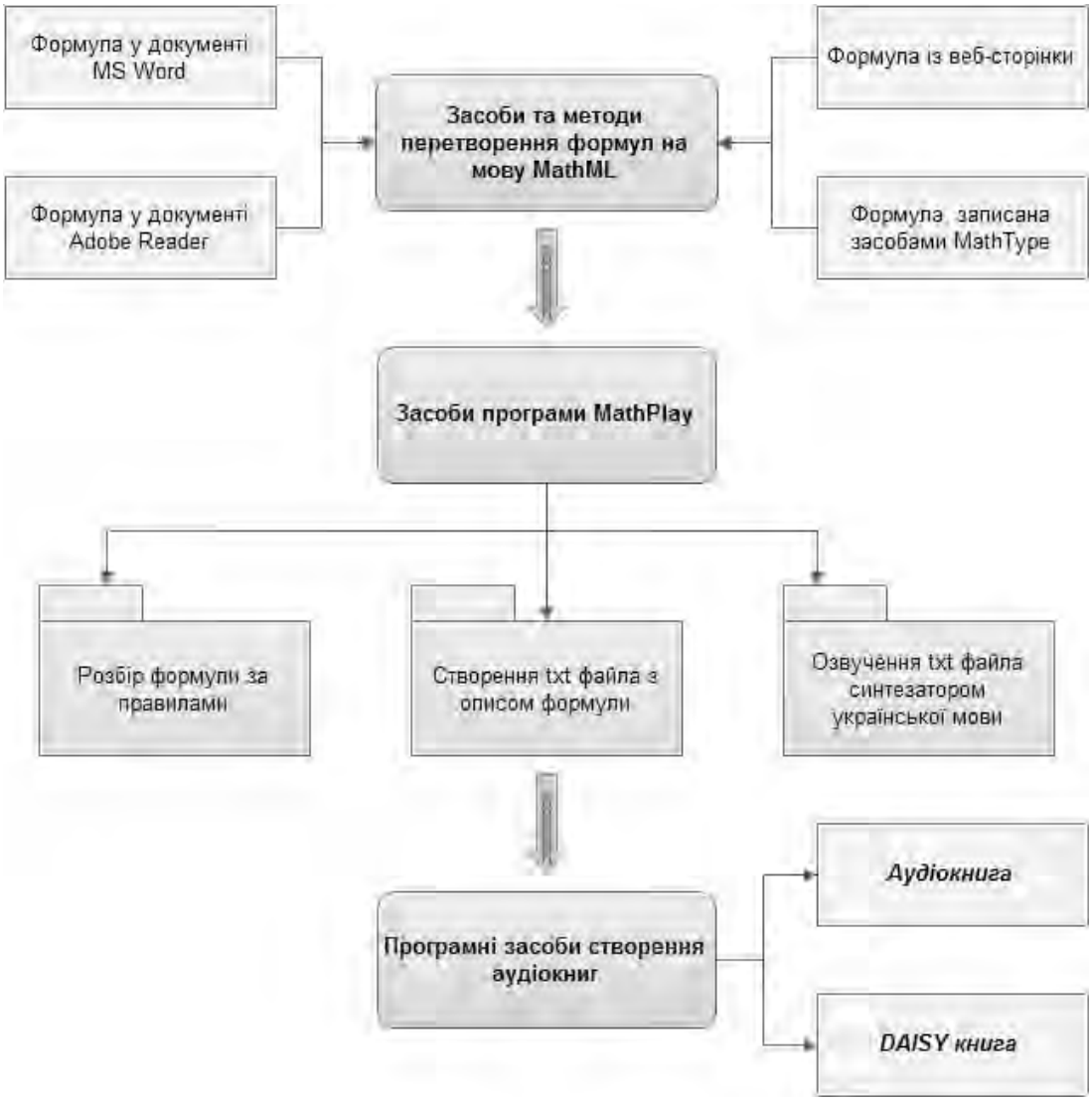


Рис. 5. Процес озвучення формул українською мовою

Висновки

Трансформаційні зміни в роботі бібліотек сприятимуть безбар'єрному доступу до інформації усіх категорій користувачів, а створення електронних бібліотек з адаптованими інформаційними ресурсами для користувачів з вадами зору підвищить ефективність їх інформаційного обслуговування. Формування масиву технічних та природничих інформаційних ресурсів електронної бібліотеки для осіб з особливими потребами українською мовою нагальна потреба сьогодення, вирішенню якої сприяє розроблена прикладна програмна система. В основу прикладної програмної системи опрацювання україномовних технічних текстів для людей з вадами зору покладено використання та удосконалення відомих методів (класифікації, експертних оцінок, трансформації синтаксичного дерева, дерев рішень, змішаних методів тощо). Розроблена прикладна програмна система забезпечує швидке визначення оптимальної структури для подання україномовного інформаційного ресурсу у DAISY форматі. Це сприяє підвищенню ефективності процесу підготовки до

перетворення у DAISY формат текстової інформації технічного та фізико-математичного спрямування, автоматизації процесу накладання навігації на книгу і дає можливість істотно скоротити часові затрати на створення DAISY книги. Досліджені та розроблені програмно-алгоритмічні засоби і методи озвучення математичних формул дали змогу озвучувати україномовні технічні тексти за допомогою синтезатора мови, забезпечили можливість наповнювати DAISY книги технічним контентом і конвертувати математичні формули у текстовий опис. Також було сформовано базу знань, у якій описано правила перетворення математичних формул із множини різних форматів на мову математичної розмітки MathML, для генерування текстового файла з описом для подальшого озвучення синтезатором української мови.

1. Automating Tactile Graphics Translation [Текст] / R. E. Ladner, M. Y. Ivory, R. Rao, S. Burgstahler, D. Comden, S. Hahn, M. Renzelmann, S. Krisnandi, M. Ramasamy, B. Slabosky, A. Martin. A. Lacenski, S. Olsen, D. Croce. // Proc. of 7th Int. ACM Sigaccess Conf. on Computers and Accessibility, January 2005, New York. – New York, 2005. – S. 50–57. 2. Automated tactile graphics translation: in the field [Текст] / Jayant C., Renzelmann M., Wen D., Krisnandi S., Ladner R., Comden D.: // Proc. of 9th Int. ACM Sigaccess Conf. on Computers and Accessibility, January 2007, New York. – New York, 2007. – S. 75–82. 3. Amick Nancy Guidelines for Design of Tactile Graphics [Електронний ресурс] / Nancy Amick, Jane Corcoran // APH Educational Research. American Printing House for the Blind, 2004, Inc. – Режим доступу: http://www.aph.org/ edresearch/guides.htm. 4. Hutchins J. The evolution of machine translation systems / W. John Hutchins // Practical experience of machine translation: Proceedings of a conference, 5–6 November 1982, UK. – London, 1982. – S. 21–37. 5. The mathematics of statistical machine translation [Текст] / Brown P. F., Pietra S. A. D., Pietra V. J. D., Mercer R. L. // Computational Linguistics. – 1993. – Vol.19(2). – S. 263–313. 6. Synchronized Multimedia Integration Language (SMIL 2.0) [Електронний ресурс] / Ayars J., Bulterman D., Cohen A., Day K., Hodge E., Hoschka P., Hyche E., Jourdan M., Kim M., Kubota K., Lanphier R., Layaida N., Michel T., Newman D., van Ossenbruggen J., Rutledge L., Saccocio B., Schmitz P., Kate W. – Режим доступу: http://www.w3.org/TR/2005/REC–SMIL2–20050107. 7. Specifications for the Digital Talking Book. ANSI/NISO Z39.86–2005 (R2012). – Режим доступу: http://www.daisy.org. 8. Лозицький О. А. Контент-керована інтелектуальна система супроводу незрячих людей / О. А. Лозицький, О. В. Пасічник // Штучний інтелект : наук.-теор. журн. – Донецьк, 2009. – № 4. – С. 437–440. 9. Handbook of Graph Grammars and Computing by Graph Transformations [Текст] / H.Ehrig, G.Engels, H.–J. Kreowski, G.Rozenberg. – Volume 2: Applications, Languages and Tools World . – Scientific, 1999. – 132 s. 10. Schwarz Brett. Customize the Speech API–for Your APP. Solutions Architect [Електронний ресурс] / Brett Schwarz. – Режим доступу: http://developer.att.com/home/community/conference/CustomizeTheSpeechAPI-ForYourApp.pdf.