



Препоръки за специфични критерии за четимост на текстовете за различните класове, свързани с лексикалното съдържание и граматическите структури

Светла Коева, Цветана Димитрова, Христина Кукова,
Валентина Стефанова, Мария Тодорова, Ивелина Стоянова

*Секция по компютърна лингвистика, Институт за български език,
Българска академия на науките*



Измерване на четимостта

- Характеристиките, използвани за измерване на четимостта на текстовете, се разделят на четири основни групи (Делорлета и др. 2011: 76–77):
 - характеристики на текста;
 - лексикални характеристики;
 - морфо-синтактични характеристики;
 - синтактични характеристики.



Измерване на четимостта

Характеристиките на текста включват:

- типа текст, дължината и структурата на текста;
- дължината на изреченията, изчислена като среден брой думи в изречение;
- дължината на думите, изчислена като среден брой букви в дума.

Лексикалните характеристики са:

- процентът на уникални думи в текста;
- процентът на лексикални единици от референтен списък спрямо възрастта;
- разпределението между „основна лексика“, думи с висока честота на употреба, думи с честа употреба и думи с относително по-ниска честота;
- броят срещания на различните думи (основна и специфична лексика) в текста.



Измерване на четимостта

Морфо-синтактичните характеристики засягат:

- съотношението на пълнозначните думи (глаголи, съществителни, прилагателни и наречия) към общия брой думи;
- използването на разнообразни словоформи.

Синтактичните характеристики включват:

- сложността на синтактичните конструкции – дълбочината на синтактичното дърво или най-дългия път от корена на дървото до листата (думите);
- изразяването или неизразяването на компонентите на простото изречение;
- средната дълбочина на синтактичното дърво на подчинените изречения;
- словореда на простите изречения;
- дължината на зависимите връзки.



Автоматичен лингвистичен анализ

- Предварителната автоматична обработка на текстовете включва:
 - разделяне на изречения;
 - токънизация;
 - тагиране (определяне на част на речта);
 - лематизация (определяне на основната форма и на граматическите характеристики);
 - идентифициране на структурата на изречението (чрез модела на универсалните зависимости).
- За българските текстове анализът е направен с Българската многокомпонентна система за първична обработка и лингвистична анотация (Koeva et al. 2020).



Автоматичен лингвистичен анализ

- Автоматичният лингвистичен анализ, извършен с помощта на съвременните инструменти за обработка на естествен език, предоставя възможност за статистически анализ и наблюдение на текстовата четимост на различни езикови равнища:
 - артикулационна сложност (дължина на думите и лексикално разнообразие);
 - покритие на различните части на речта и техните словоформи;
 - лексикална типичност (статистическа честота на употребата и разпространението на думите);
 - морфемен състав на думите;
 - синтактични функции на думите в изреченията;
 - словоредни позиции и съчетаемост на думите.



Индекси за оценка на четимост

Съществуват редица индекси за измерване на четимостта на текста въз основа на редица характеристики.

Индекс на Флеш (Flesch Reading Ease; от 100 / лесен до 0 / труден):

$$FLI = 206.835 - 1.015 \times (\text{средна дължина на изречение}) - 84.6 \times (\text{средна дължина на дума})$$

Ликс индекс (LIX, Läsbarhetsindex; скала от 20 / лесен до 60 / труден; дълги думи са тези с над 6 символа):

$$LIX = (\# \text{ думи} / \# \text{ изречения}) + (\text{дълги думи} \times 100) / \# \text{ думи}$$

Индекс на Колман и Лиау (Coleman-Liau Index; дава приблизително ниво на трудност като брой години училищно образование):

$$CLI = 0.0588 \times (\# \text{ букви} / \# \text{ думи}) \times 100 - 0.296 \times (\# \text{ думи} / \# \text{ изречения}) \times 100 - 15.8$$



Индекси за оценка на четимост

Щом се здрачи, рисът, който спеше на топло в хралупата на един грамаден бор, се измъкна мързеливо навън. Той се прозя с широка, котешка гримаса, от която неговите уши се свиха назад, а раззиналата му уста се показаха белите му, остри зъби. Наоколо беше тихо и мрачно. По напуканите стъбла на тъжните стари борове висяха черни мъхове. Шишарки и клечки, сплетени като гарванови гнезда, се тъмнееха в клоните им, където се белееха и остатъци от вледенен сняг.

Емилиян Станев, „Рис“ (разказ)

FLI = 21

мн. труден (колеж)

LIX = 41

труден (10. клас)

CLI = 11.5

труден (12. клас)



Индекси за оценка на четимост

- Установяваме съотнесимост на оценките на четимостта, но няма пълно припокриване.
- Индексите са създадени и тествани за други езици (английски, шведски); изискват адаптиране за български и за славянските езици (напр. за руски се използват формулите с коригирани константи, установени емпирично).
- Оценката на текстовете чрез разгледаните мерки е непълна, тъй като не се вземат предвид лексикалните, словообразователните и морфосемантичните специфики на думите за български език, нито комплексът от семантични, селективни и граматични ограничения, произтичащи от българския словоред, лексикална съчетаемост и синтактични конструкции.
- По тези причини е необходимо да се създаде специфична за българския език оценка на четимостта на текстове.



Автоматичен анализ на четимостта на думите в текста

Цел: Да се създаде (специфичен за българския език) модел за автоматична класификация на думите и по-големите езикови единици (фрази, изречения) според трудността на четене.

Задачи:

- Да се установи кои формални характеристики на думите са **значими** при изграждането на модела.
- Да се анализират **зависимостите** между трудността на четене на отделните думи и последователностите от думи.
- Да се разработи система от **критерии за оценка** на четивната трудност, които могат да бъдат приложени в модел за автоматична класификация на текстовете.



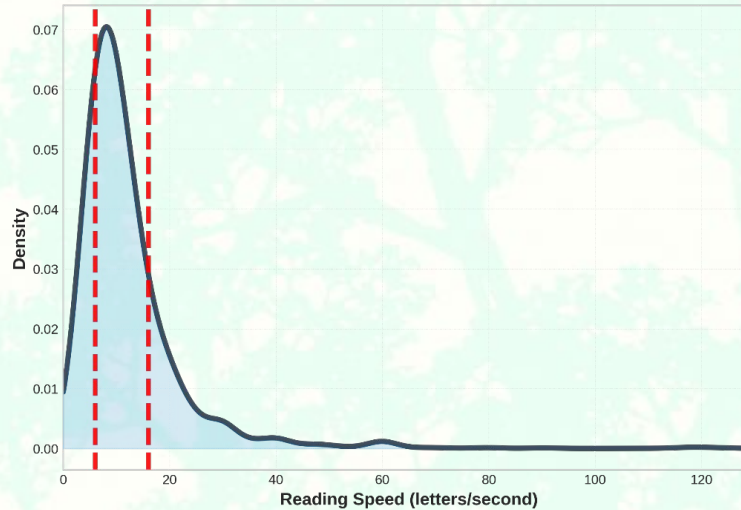
Автоматичен анализ на четимостта на думите в текста

- **Мярка за трудност:** Средна скорост на четене за думата, изчислена като брой прочетени букви в секунда.
- Думите се класифицират като: **трудни / бавни при прочит** (прочетени с бавна скорост), **средни** (прочетени с най-срещаната скорост) и **лесни / бързи при прочит** (прочетени с висока скорост).
- Данните за скорост на четене не са пропорционално разпределени (налице са голям брой различни лесни думи с малка честота). Определянето на категориите става въз основа на разпределението – за **нормална** се приема скоростта, отбелязана при най-много думи (60%); лявата опашка са **трудни** (20%), а дясната опашка са **лесни** (20%).

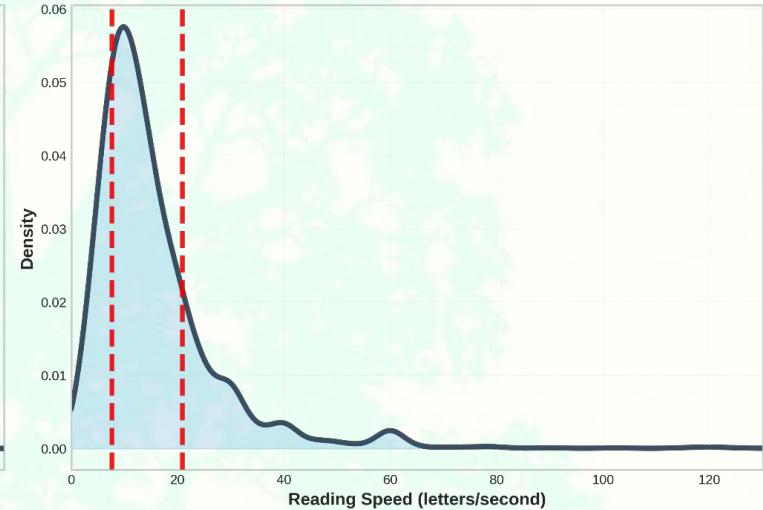


Автоматичен анализ на четимостта на думите в текста

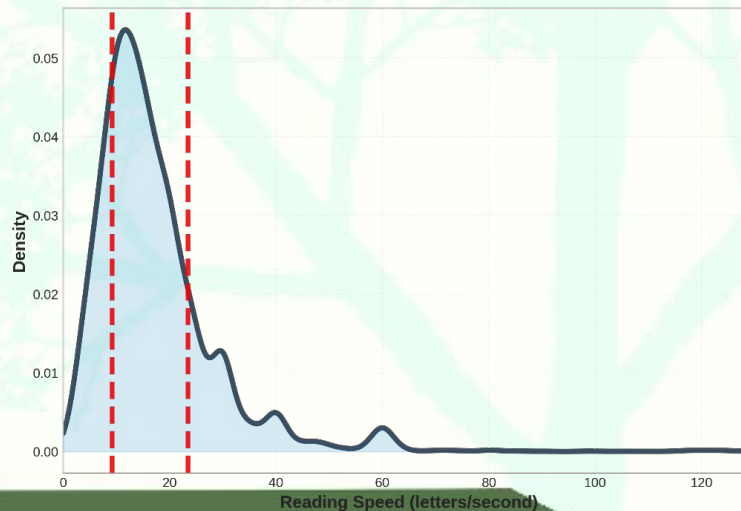
Grade 2



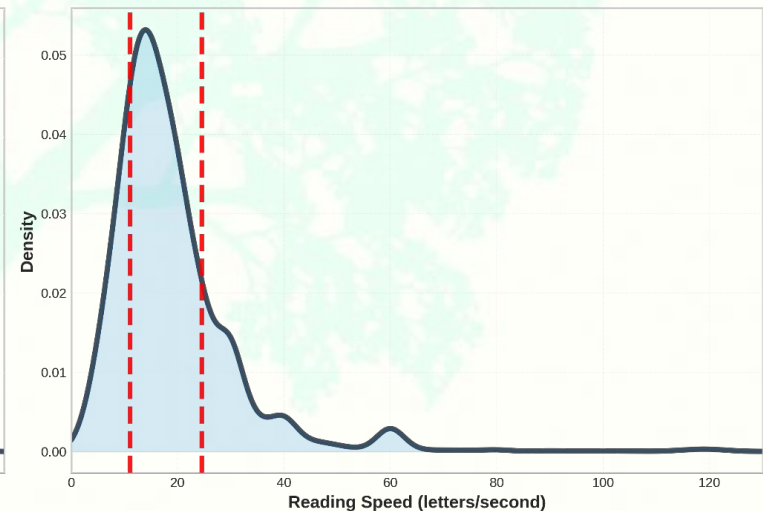
Grade 3



Grade 4



Grade 5





Автоматичен анализ на четимостта на думите в текста

Клас	Бавни (трудни)	Средни	Бързи (лесни)
2	< 6.11	6.11 - 14.04	≥ 14.04
3	< 7.73	7.73 - 18.69	≥ 18.69
4	< 9.24	9.24 - 20.08	≥ 20.08
5	< 11.19	11.19 - 22.34	≥ 22.34



Автоматичен анализ на четимостта на думите в текста

Най-често срещани трудни думи:

2. клас: *къща, внушителни, стичат, върне, набраздена, зали, поздрав, приятния*

3. клас: *небосвода, надявайки се, мистериозни, страст, кула, сълзи, пръчка*

4. клас: *риск, привлече, снощната, обзаведена, прелетя, печена, авторитета, истинско*

5. клас: *монолог, извини, обсъждат, месоядно, изтъркана, мистериозни, ентусиазъм, отстрани*

Многосрични думи и думи със струпване на съгласни.



Автоматичен анализ на четимостта на думите в текста

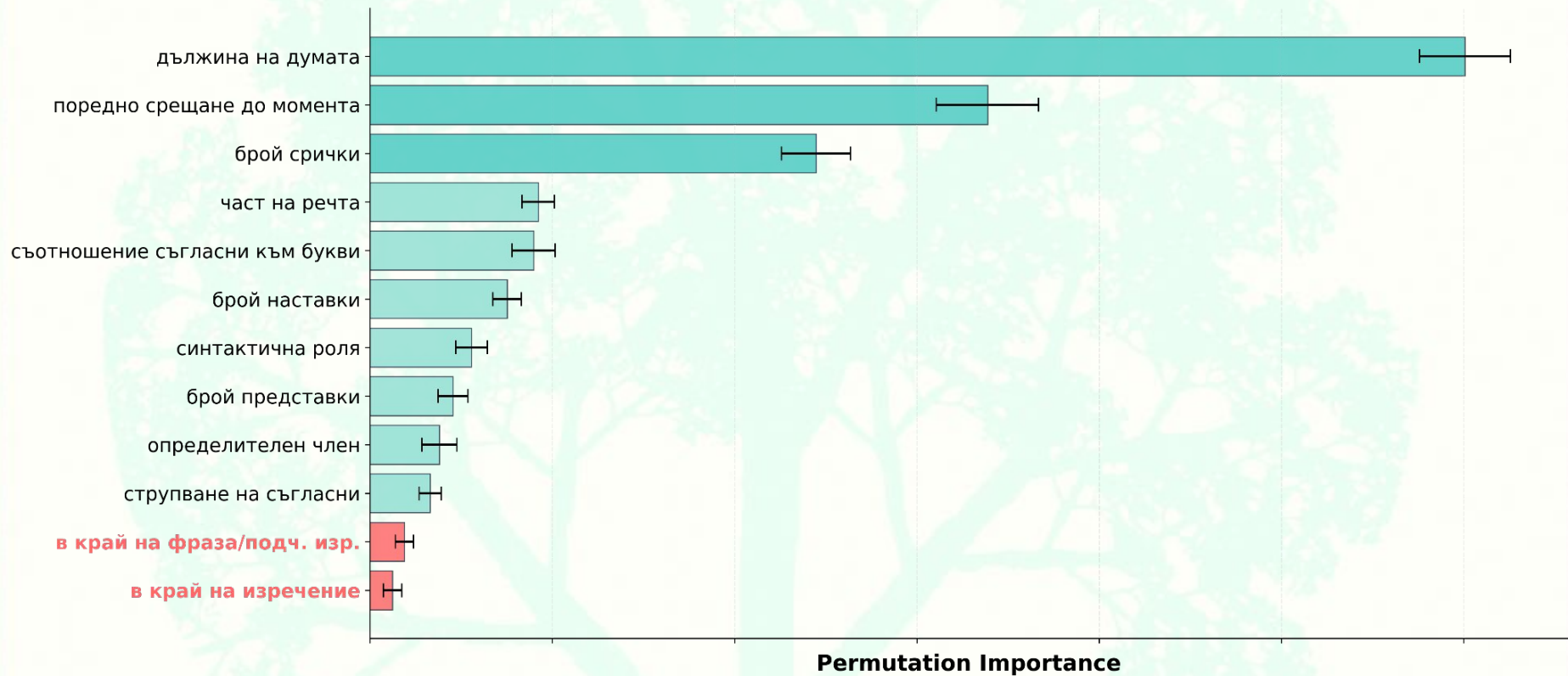
Метод: Класификатор на Java, използващ машинно обучение чрез многослойни перцептрони (Multilayer Perceptron, MLP)

12 характеристики (фонетични, морфологични, морфосинтактични, синтактични, текстови) за машинното обучение:

дължина на думата в брой букви; дължина в брой срички; съотношение съгласни към общ брой букви в думата; брой струпвания на 3+ съгласни; част на речта; синтактична функция; брой представки; брой наставки; наличие на определен член; брой срещания в текста до момента; позиция на думата спрямо граница на фраза или изречение.



Автоматичен анализ на четимостта на думите в текста





Автоматичен анализ на четимостта на думите в текста

Значимост на отделните характеристики за MLP

- Всички изброени характеристики имат положителен принос, т.е. имат значение за определянето на трудността на четене на думите, но с различна тежест.
- По-внимателният анализ показва, че:
 - само характеристиките на думата не са достатъчни за определяне на четимостта, измерена като скорост на четене; необходимо е да се анализира и контекстът;
 - скоростта на четене не е абсолютен показател за качеството на четенето, защото някои случаи на забавяне при четене са нормални, очаквани и показват осмисляне на текста.



Наблюдения

1. Някои бавни думи се дължат на нормално учленяване на смисловите единици в текста (пауза в края на фраза, просто изречение или изречение):

- В края на изречение 22.2% от думите са бавни за прочит; 64.7% средни; само 13.1% са бързи.
- Скоростта на прочитане на същите думи в край на изречение спрямо други позиции е 71% по-висока.
- На границата на фраза или просто изречение (маркирана със запетая или съюз) 29.7% са бавни; 55.5% са средни; само 14.8% са бързи.
- Преди запетая или съюз същите думи са средно с 64% по-бавни спрямо други позиции в изречението.



Наблюдения



2. Местоименията са кратки, познати, високочестотни, но с относително бавен прочит (предполага се, че изискват допълнително време за декодиране на референциалността): *си, ми, му, го, ме* (както и *някои, които*):

- 18.5% от местоименията са бавни;
- 13% от всички бавно прочетени думи са местоимения.



Наблюдения

За местоименията:

- Най-бавни са комбинациите същ. + възвр. прит. местоимение: *приятелите си, майка си, полета си*;
- Следват останалите притежателни местоимения за всички лица: *дядо ми, майка му, погледът му, листенцата му*;
- Бавни са комбинации от две последователни местоимения: *ми го, себе си, тя ме, тя ги, мен ми, който я*;
- Бавни са и *който, които*, въвеждащи подчинено изречение.



Наблюдения



Конструкции с кратки притежателни местоимения

Клас	Среден брой конструкции същ. + прит. местоимение в изречение	% от тях бавни
2	0.252	16.8%
3	0.316	18.6%
4	0.319	19.1%
5	0.341	22.1%
средно	0.307	19.2%



Наблюдения

Последователност от две местоимения

Клас	И двете бързи	1 бързо + 1 норм.	И двете норм.	1 бавно + 1 друго	1 друго + 1 бавно	И двете бавни
2	2.3%	15.9%	43.2%	11.4%	15.9%	11.4%
3	11.1%	17.4%	36.1%	12.5%	11.8%	11.1%
4	11.6%	23.7%	31.6%	11.6%	11.6%	10.0%
5	9.2%	14.5%	39.5%	14.0%	10.5%	12.3%
средно	8.5%	17.9%	37.6%	12.4%	12.5%	11.2%

36% бавни



Наблюдения

3. Има редица високочестотни кратки глаголи, които са **бавни**: *е, беше, бил, има, знае, може*:

- 19.4% от късите глаголи са бавни;
- 6% от всички бавно прочетени думи са глаголи с дължина до 4 букви.



Наблюдения

За спомагателните глаголи:

- В началото на изречение (когато изречението започва с глагол) 28.5% от спомагателните глаголи са **бавни**.
- По-бавното прочитане е характерно за глаголите изобщо в абсолютното начало на изречението – 30.0% от глаголите в началото на изречение са **бавни**.
- При личния глагол *имам* явлението е още по-често срещано – 34.2% от употребите на *имам* в началото на изречение са **бавни**.
- За сравнение, средно само 17.8% от думите в началото на изречение са бавни.



Автоматичен анализ на четимостта на думите в текста

Глагол в началото на изречение в сравнение с дума в началото на изречение

Клас	% бавни глаголи	% бавни глаголи в нач. на изр.	% бавни спом. глаголи	% бавни спом. глаголи в нач. на изр.	% бавни думи	% бавни думи в нач. на изр.
2	22.1%	30.9%	16.7%	24.1%	13.5%	18.8%
3	20.6%	28.1%	18.8%	27.4%	13.8%	17.6%
4	21.9%	31.0%	18.6%	31.9%	13.8%	16.8%
5	20.4%	28.7%	20.7%	30.7%	13.7%	18.0%
Средно	21.3%	30.0%	18.7%	28.5%	13.7%	17.8%



Автоматичен анализ на четимостта на изречението

1. Ключовата характеристика, определяща скоростта на четене на ниво изречение, е лексикалното съдържание: пропорционалното участие на **трудни** думи и на **лесни** думи в изречението.
2. На следващо място е броят срещания (до момента) на думите в текста. Повече думи с по-висока честота и познати думи правят четимостта по-добра.
3. Мерките за четимост стават по-достоверни за по-големите езикови единици – на ниво изречение, отколкото на ниво дума.
4. За оценка на четимостта на ниво изречение се изисква обаче **достоверна оценка** за четимостта на ниво дума.



Препоръки

- Да се обръща внимание преди или след четене на непознати или по-трудни за усвояване думи и на такива, които изискват повече време за четене.
- В текстовете новите и непознатите думи, които трябва да се усвоят, да се употребяват повече от веднъж, в различна форма и в различна синтактична позиция, защото различните форми се усвояват различно (по-сложните морфологични форми се четат по-бавно).



Препоръки

- Да се съчетава четенето с упражнения върху използването и разбирането на местоименията (правилно разпознаване на антецедента в по-сложни случаи).
- Специално внимание да се обръща на употребата на възвратните местоимения – притежателно и лично, както и на правилното разпознаване на антецедентите от контекста.



Препоръки

- Да се обръща внимание на изречения, които започват с глагол, особено такива със сложни глаголни форми, при които в началото на изречението стои спомагателен глагол.
- Специално внимание да се обръща на сегментирането на текста при четене и паузите при граница на фраза и изречение.



Благодарим за вниманието!

Проектът се финансира от Българската академия на науките
и Националния съвет за научни изследвания в Италия.

Научете повече на страницата на проекта!

dcl.bas.bg/literacy/

