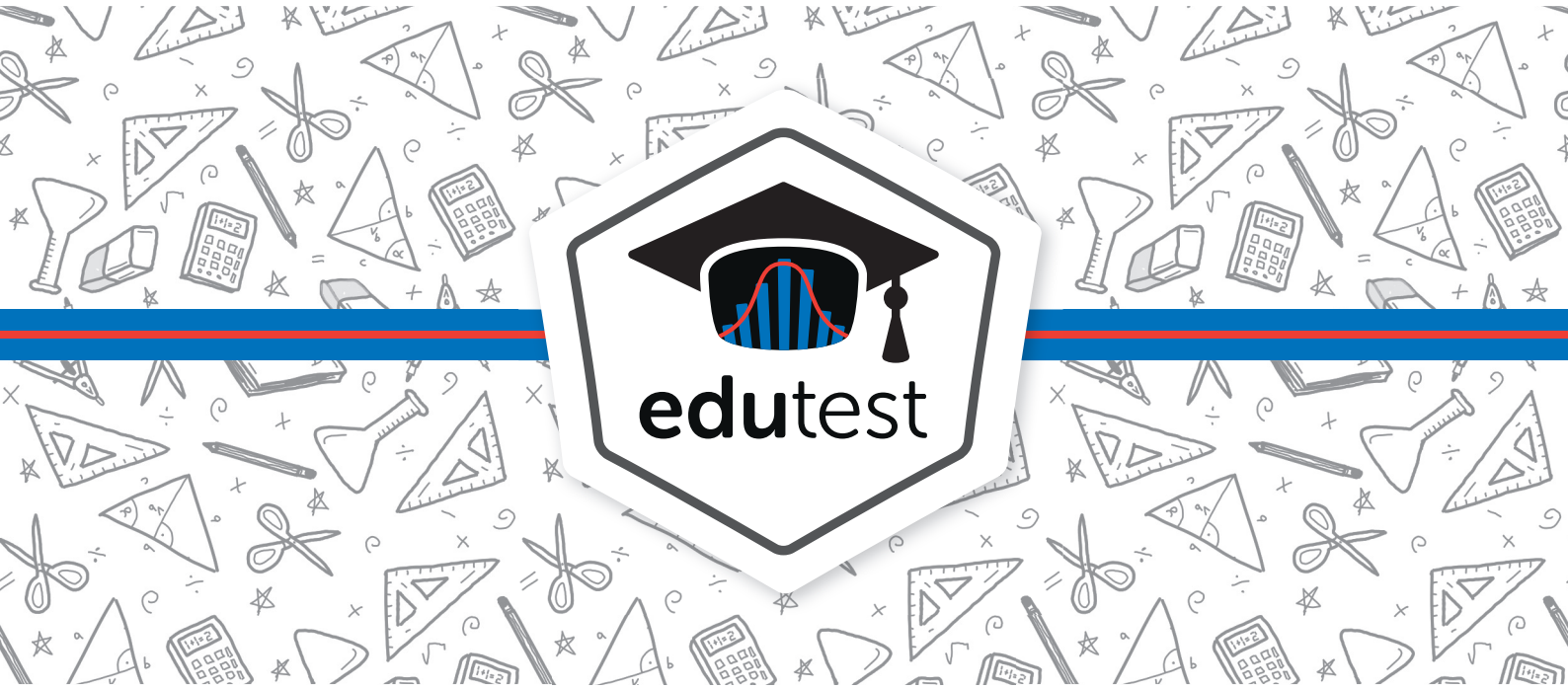


Soubor postupů a nástrojů
pro zkvalitnění tvorby znalostních testů pomocí psychometrických modelů

Patrícia Martinková, Eva Potužníková, Jan Netík a kol.

PEDAGOGICKÁ FAKULTA
Ústav výzkumu a rozvoje vzdělávání
Univerzita Karlova

Ústav informatiky
Akademie věd ČR

T A
Č R

Projekt TL05000008 Výzvy pro hodnocení znalostí: Analytická podpora tvorby znalostních testů byl spolufinancován se státní podporou Technologické agentury ČR v rámci Programu ÉTA 5.

Obsah

Predikce obtížnosti položek pomocí modulu EduTest Text Analysis

Jana Dlouhá, Jan Netík, Lubomír Štěpánek,
Eva Potužníková, Patrícia Martinková

1 Úvod

Odhad obtížnosti testových položek je klíčový pro vytváření spravedlivých a spolehlivých testů, které efektivně měří znalosti a dovednosti. Tyto odhady se obvykle zakládají na datech z pilotního testování nebo na hodnocení odborníků. Když je pilotní testování prováděno na malých vzorcích respondentů nebo není možné, nabývají na významu expertní odhady obtížnosti položek. V poslední době se ukazuje rostoucí tendence kombinovat tyto expertní odhady s algoritmy strojového učení, což může výrazně zvýšit jejich přesnost (Alkhuzaey & Tendeiro, 2020).

Modul *EduTest Text Analysis* pro predikci obtížnosti položek pomocí textové analýzy vyvinutý v rámci projektu EduTest vychází z našeho teoretického výzkumu v oblasti využití algoritmů strojového učení a textové analýzy položek (Štěpánek et al., 2023). Modul zpřístupňuje námi navrženou metodu analýzy textu pomocí algoritmů strojového učení prostřednictvím jednoduchého interaktivního rozhraní, do něhož uživatel pouze vloží zadání testové položky a spustí analýzu.

Výstupem z analýzy jsou tabulky obsahující řadu textových charakteristik, které mohou potenciálně zvyšovat nebo naopak snižovat obtížnost analyzované položky. Tvůrce testu je může přehledně vyhodnocovat a porovnávat dosažené hodnoty pro nově vytvářenou položku s hodnotami položek uložených v databázi modulu. Z vybraných textových charakteristik modul dále vypočítá odhadovanou obtížnost analyzované položky pomocí prediktivního modelu navrženého na základě známých charakteristik již administrovaných testových položek.

Prediktivní model vychází z textových vlastností dříve zadávaných testových položek a z jejich obtížnosti odhadnuté z žákovských odpovědí. Tato data byla využita k trénování algoritmů strojového učení, které se z obrovského množství informací o textovém znění zadaných položek snaží určit, které textové charakteristiky mohou nejlépe predikovat jejich obtížnost, která je známá, resp. zjistitelná z empirických žákovských odpovědí. Natrénované algoritmy byly následně aplikovány na predikci obtížnosti nových položek a z mnoha potenciálních prediktivních modelů byl vybrán ten, který obtížnost položek predikoval nejpřesněji.

Modul *EduTest Text Analysis* je součástí širší aplikace Shiny Item Analysis (SIA) a je dostupný pod záložkou Modules na odkazu

https://shiny.cs.cas.cz/ShinyItemAnalysisEduTest/

Předpokládané využití modulu je především pro doplnění a zpřesnění expertních odhadů tvůrců testů na nových položkách. Zároveň může modul sloužit k demonstraci příslušných metod textové analýzy a strojového učení.

Modul byl vyvíjen ve spolupráci s Centrem pro zjišťování výsledků vzdělávání (CZVV) a v současné době je přizpůsoben pro vybrané testové položky z maturitních testů z anglického jazyka zaměřené na hodnocení porozumění textu. Ve stávající verzi modul obsahuje databázi položek z části 5 maturitních testů z anglického jazyka zadávaných v letech 2016–2023. Jedná se o položky složené z relativně krátkého výchozího textu (v rozsahu jednoho odstavce), otázky a čtyř nabízených možností odpovědi, z nichž právě jedna je správná. Předpokládaná úroveň jazykových znalostí hodnocených těmito položkami je úroveň B1 dle Společného evropského referenčního rámce pro jazyky (CEFR). Algoritmy implementované v modulu tedy odhadují obtížnost položek na základě textových charakteristik, které nejlépe fungují právě pro tento typ položek.

2 Příprava textu pro analýzu

Pro vlastní analýzu je nutné texty upravit tak, aby byly vhodné pro strojové zpracování. Modul *EduTest Text Analysis* využívá pro přípravu textů standardní metody, které jsou podrobně

popsány například v publikaci Hvitfeldta a Silgeové o analýze textů pro malé až středně velké datové sady (Hvitfeldt & Silge, 2021). Příprava textu pro strojové zpracování probíhá v modulu automatizovaně. Procesy přípravy textu implementované v modulu kladou důraz na vyváženost mezi důkladným strukturováním vstupních textových dat a zachováním klíčových informací. Tento přístup je nezbytný pro zajištění přesné a hloubkové analýzy textu, která poskytne smysluplné interpretace a závěry. Hlavními kroky v přípravě textu jsou tzv. tokenizace, lemmatizace, odstranění informačně chudých slov a čištění textu, které si nyní představíme blíže.

Rozklad textu na segmenty (tokenizace) Původní souvislý text je rozdělen na menší části, jako jsou věty, slova nebo znaky. Toto dělení umožňuje detailnější zkoumání každého segmentu, či „tokenu“, což jsou jednotlivé stavební prvky textu.

Úprava slovních tvarů (lemmatizace) Lemma je reprezentativní podoba slova uváděná ve slovnících. V rámci lemmatizace se slova upraví na jejich základní (slovníkový) tvar, například slova „psala“, „psali“ a „psát“ se v rámci lemmatizace všechna převedou na základní tvar „psát“. Na rozdíl od stematizace, která spočívá v odstranění předpon a přípon a převedení slov na jejich kmen, jako například „psa“, lemmatizace udržuje plný význam slov.

Odstranění informačně chudých slov (tzv. stop slov) V textech se běžně vyskytuje množství často používaných slov, která sama o sobě nenesou žádný význam (tzv. stop slov). Jsou to například předložky, spojky, zájmena, pomocná slovesa nebo členy, tedy zpravidla velmi krátká slova, která nepřinášejí užitečné informace o textu a kvůli své vysoké frekvenci mohou zkreslovat jeho analýzu. Vypuštění stop slov neovlivňuje srozumitelnost a význam sdělení a umožňuje soustředit se na podstatné informace v textu. Pro eliminaci stop slov jsou v jednotlivých jazycích k dispozici jejich seznamy. Modul využívá seznam SMART, který zahrnuje 174 takových slov (Hvitfeldt & Silge, 2021; Wilbur & Sirotkin, 1992).

Čištění a zpřesnění textu Konečným krokem je odstranění dalších nepotřebných znaků a slov, která nenesou význam. V testových položkách z anglického jazyka jsou to například poznámky pod čarou s českými překlady méně známých anglických slov použitých ve výchozím textu, HTML značky v údajích o zdroji výchozího textu, bílé znaky a speciální znaky. Díky tomu je možné se lépe zaměřit na to nejdůležitější v textu.

3 Textové charakteristiky položek

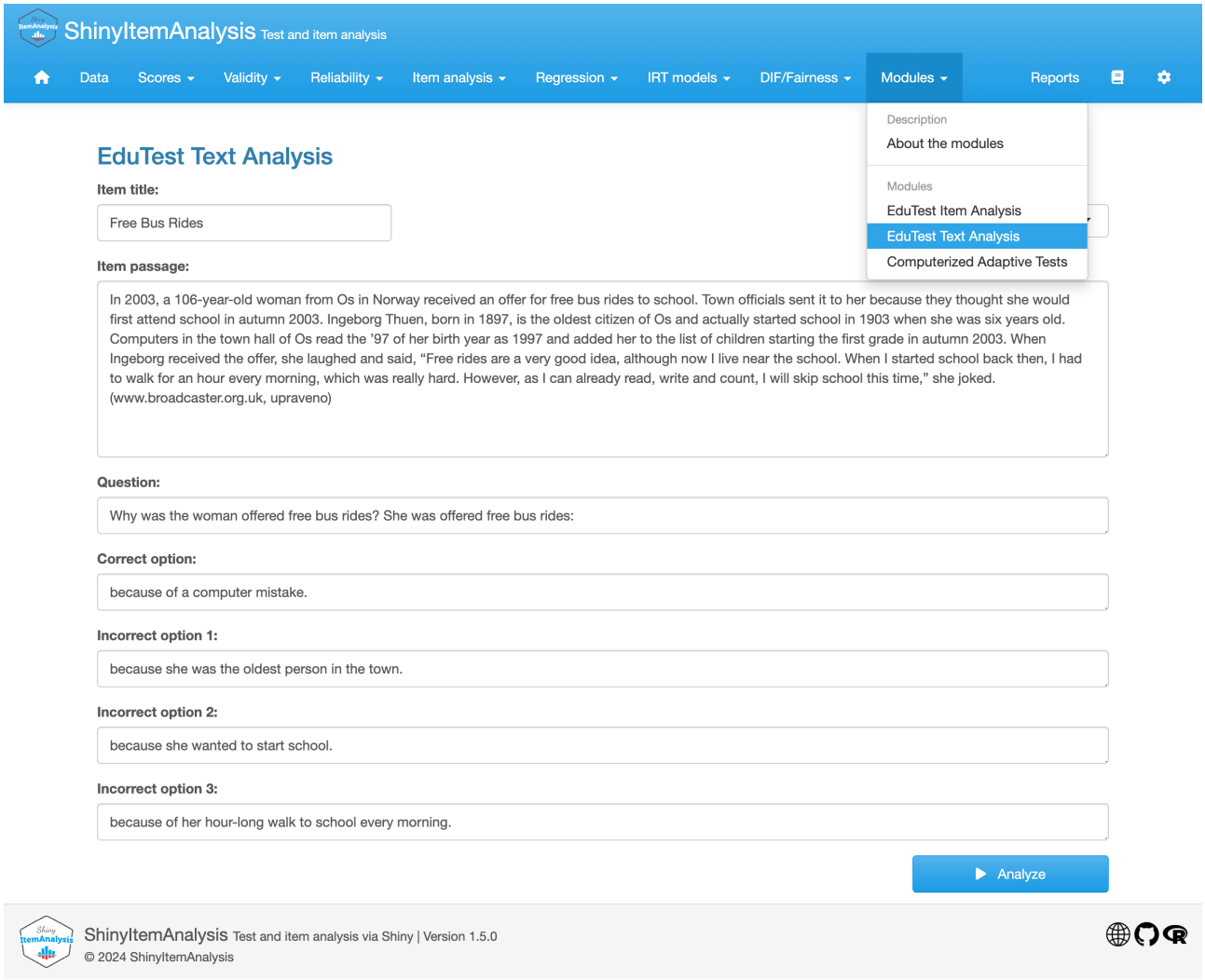
V rámci textové analýzy testových položek je možné z připravených textových dat extrahovat různé vlastnosti textu. Modul *EduTest Text Analysis* vyhodnocuje čtyři hlavní typy textových vlastností, které mohou zvyšovat nebo naopak snižovat obtížnost položek – základní charakteristiky textu, kategorizaci v textu použitých slov z hlediska frekvence jejich užívání v jazyce, celkovou míru čitelnosti textu a míru podobnosti mezi různými částmi testové položky.

V modulu jsou zahrnuty textové charakteristiky, které byly analyzovány ve studii Štěpánka a kol. (2023). Tyto charakteristiky byly zvoleny na základě dosavadních výzkumných zjištění (Alkhuzaey & Tendeiro, 2020; Beinborn et al., 2014, 2015; Ferrara, 2022), která ukázala, že obtížnější jsou zpravidla položky obsahující větší množství textu, delší slova a položky využívající celkově složitější jazyk charakterizovaný větším výskytem méně frekventovaných slov a nižší mírou čitelnosti. Vyšší obtížnost lze také očekávat u položek, jejichž distraktory (nesprávné možnosti odpovědi) jsou podobné s výchozím textem a se správnou odpovědí, a u těch položek, jejichž správné odpovědi jsou málo podobné s výchozím textem. Takové položky vyžadují větší mentální úsilí pro pochopení a výběr správné odpovědi.

3.1 Základní charakteristiky textu a odvozené ukazatele

Nejjednodušší vlastnosti, jimiž lze charakterizovat textová zadání testových úloh, se týkají lexikální a morfologické analýzy na úrovni znaků a slov. Patří sem například počet znaků, slabik, slov a tokenů. Z těchto údajů lze dále vypočítat odvozené hodnoty, jako je průměrná délka slov ve znacích apod.

Zmíněné vlastnosti jsou měřeny pro celý text položky, tj. dohromady pro znění výchozího textu, otázky a jednotlivých nabízených odpovědí (správné odpovědi i distraktorů). Aby bylo možné rozlišit efekt vlastností odvozených z různých částí testové položky (výchozího textu, otázky, správné odpovědi, distraktorů), měříme je také pro každou část samostatně. Jednotlivé části položky je potřeba specifikovat už při jejím zadávání do vstupního rozhraní modulu pro analýzu textu (obrázek 1).[1]



Obrázek 1: Zadání položky v modulu *EduTest Text Analysis*

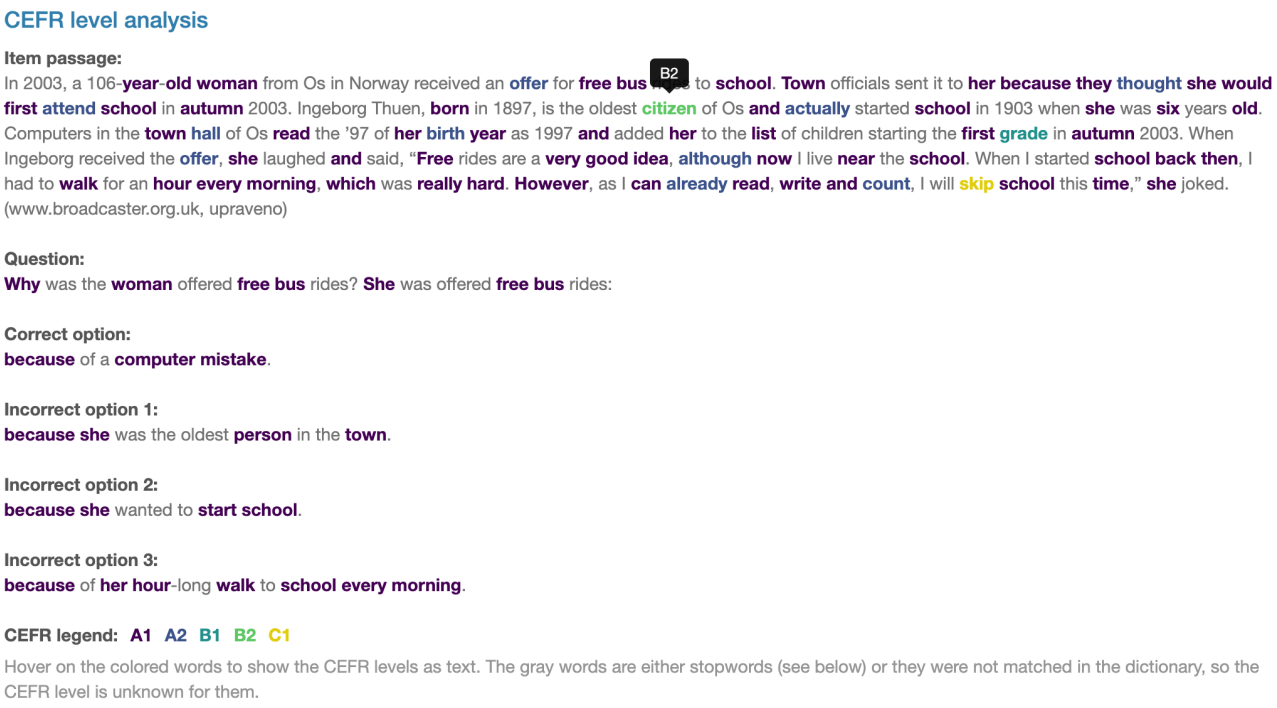
Po spuštění analýzy pomocí tlačítka **Analyze** se základní charakteristiky textu a odvozené ukazatele zobrazí ve výstupu z analýzy formou přehledné tabulky, společně s dalšími textovými charakteristikami, které si popíšeme v následujících částech.

[1] V této ukázce vkládáme zadání jedné z položek zveřejněné na https://maturita.cermat.cz/menu/testy-a-zadani-z-predchozich-obdobi/anglicky-jazyk/testy-a-zadani-anglicky-jazyk.

3.2 Kategorizace slov dle frekvence jejich užívání v jazyce

V souladu s předpokladem, že zařazení méně běžných slov zvyšuje obtížnost testových úloh, jsou v modulu implementovány také metody, které vyhodnocují frekvenci výskytu slov. K vyhodnocování frekvence výskytu slov lze zvolit různé přístupy.

Při hodnocení znalosti cizích jazyků se často zohledňují úrovně podle Společného evropského referenčního rámce pro jazyky (CEFR). Proto jsou v modulu implementovány metody pro klasifikaci slov dle úrovní CEFR, které do nižších úrovní řadí běžně používaná slova a do vyšších úrovní méně běžná slova včetně specializovaných odborných termínů. Konkrétně se zde vychází ze standardů Oxford 3000™ obsahujícího 3 000 slov z úrovní A1 až B2 a Oxford 5000™ obsahujícího dalších 2 000 slov z úrovní B2 až C1. Klasifikace slov podle úrovní CEFR může nejen přispět k zpřesnění odhadu obtížnosti testových položek, ale také může tvůrcům testů pomoci upravit zadání položek tak, aby odpovídalo předpokládané úrovni jazykových znalostí, kterou má test ověřovat (např. maturitní test z anglického jazyka má ověřovat, zda žák dosáhl jazykové referenční úrovně B1). Klasifikace jednotlivých slov podle úrovní CEFR je zpřístupněna v interaktivním modulu v horní části výstupu. Po vložení textu jednotlivých částí položky a odkliknutí tlačítka **Analyze**, jak bylo popsáno výše, se barevně zvýrazní slova výchozího textu dle jejich CEFR úrovně (obrázek 2).



Obrázek 2: Výstup CEFR analýzy v modulu *EduTest Text Analysis* pro položku 25 z jarního termínu 2017

Jiný přístup k vyhodnocování frekvence užívání slov v jazyce představují jazykové korpusy. Ve studii Štěpánka a kol. (2023) jsme pracovali s klasifikací slov na velmi běžná, běžná, méně obvyklá, vzácná a velmi vzácná, která vychází z veřejně přístupného seznamu nejčastějších slov z Korpusu současné americké angličtiny (COCA) (Davies, 2008, 2011). Tento seznam obsahuje 5 000 anglických slov (lemmat) spolu s informacemi o jejich četnostech v 8 různých typech textů (beletrie, zpravodajství, internetové blogy, titulky k filmům, akademické texty atd.) a o jejich slovním druhu. Slova, která nejsou zahrnuta v tomto seznamu, jsou klasifikována jako neidentifikovaná a hodnocena jako méně běžná než ta kategorizovaná jako velmi vzácná. Klasifikace slov dle korpusu COCA zatím není součástí modulu *EduTest Text Analysis*, v budoucnu může být přidána pro rozšíření možností textové analýzy.

3.3 Hodnocení obtížnosti textu: indexy čitelnosti

Pro vyčíslení, jak obtížné může být pro čtenáře porozumět textu položky, lze dále použít tzv. indexy čitelnosti. V modulu jsou dostupné čtyři z nich: index Dalea a Challové, index Tränkla a Bailera, Gunningův Fog index a index SMOG (Simple Measure of Gobbledygook) (Chall & Dale, 1995; Kincaid et al., 1975; McLaughlin, 1969; Tränkle & Bailer, 1984). Každý index má svůj vlastní vzorec pro výpočet čitelnosti, který zohledňuje faktory jako délku slov a vět, složitost slov a počet určitých typů slov, jako jsou předložky nebo spojky.

1. **Index Dalea a Challové**: Tento index bere v úvahu délku vět a počet obtížných slov v textu. Vyjadřuje čitelnost textu jako počet let vzdělání podle amerického systému potřebných k porozumění textu. Čitelnost je vypočítána podle vzorce:

$$64 - \left(0{,}95 \times 100 \times \frac{n_{dw}}{n_w}\right) - (0{,}69 \times avg_{sl}), \tag{1}$$

   kde n_{dw} je počet slov, která nejsou zahrnuta v Dale-Challové seznamu známých slov, n_w je celkový počet slov a avg_{sl} je hodnota vypočtená jako podíl počtu slov a počtu vět (Chall & Dale, 1995).

2. **Index Tränkla a Bailera**: Je podobný indexu Dalea a Challové, ale zohledňuje navíc počet znaků ve slovech. Tento index je obzvláště užitečný pro hodnocení textů pro mladší čtenáře. Kombinuje několik faktorů: průměrnou délku slov a vět, počet předložek a v některých variantách dokonce počet spojek (Tränkle & Bailer, 1984). Čitelnost textu je vypočítána podle vzorce:

$$224{,}68 - \left(79{,}83 \times \frac{n_c}{n_w}\right) - \left(12{,}24 \times \frac{n_w}{n_s}\right) - \left(129{,}29 \times \frac{n_p}{n_w}\right), \tag{2}$$

   kde n_w je celkový počet slov v textu, n_c je počet znaků, n_s je počet vět a n_p je počet předložek v textu. Hodnoty indexu bývají záporné; čím méně záporné (blíže k nule), tím je text jednodušší, naopak velmi záporné hodnoty naznačují vyšší náročnost textu.

3. **Gunningův Fog index**: Měří složitost textu na základě délky vět a průměrného počtu slabik ve slově. Je vhodnější pro texty určené starším čtenářům. Pro odhad složitosti textu zohledňuje poměr kratších slov a komplexních slov (s třemi a více slabikami). Čím vyšší je hodnota indexu, tím pokročilejší vzdělání čtenáře je potřeba k porozumění textu (Kincaid et al., 1975). Čitelnost textu je vypočítána podle vzorce:

$$\left(\frac{n_{sw} + 3 \times n_{mw}}{100 \times \frac{n_s}{n_w}} - 3\right)/2, \tag{3}$$

   kde n_{sw} je počet krátkých slov v textu, n_{mw} je počet víceslabičných slov (slov s třemi a více slabikami), n_s je počet vět a n_w je celkový počet slov v textu. Fog index se pohybuje v rozmezí 0–20 a lze jej interpretovat jako počet let vzdělání potřebných k porozumění hodnocenému textu, přičemž hodnoty nad 13 znamenají požadavek alespoň středoškolského vzdělání a hodnoty nad 17 předpokládají alespoň vysokoškolské vzdělání, aby byl čtenář schopen textu porozumět.

4. **SMOG index**: Index SMOG (The Simple Measure of Gobbledygook) se zaměřuje na délku vět a počet složitých, víceslabičných slov. Spočítá čitelnost na základě počtu složitých slov a celkového počtu vět (McLaughlin, 1969). Vyšší počet víceslabičných slov

obvykle znamená, že text je složitější. Stejně jako Fog index je vhodnější pro texty určené starším čtenářům. Využívá následující vzorec:

$$1{,}043 \times \sqrt{n_{mw}} \times \frac{30}{n_s} + 3{,}1291, \tag{4}$$

kde n_{mw} je počet víceslabičných slov.

Vypočtené hodnoty indexů pro celý text položky i její jednotlivé části jsou v modulu ve výstupu z analýzy zobrazeny v tabulce spolu se základními vlastnostmi textu (obrázek 3). Kromě dosažených hodnot jednotlivých charakteristik je v rozbalovacím políčku uvedeno také rozpětí hodnot dané textové charakteristiky v položkách z databáze modulu (tj. v položkách z části 5 maturitních testů z anglického jazyka zadávaných v letech 2016–2023), medián a hodnota percentilu pro analyzovanou položku. Díky hodnotě percentilu lze poznat, zda je hodnota dané charakteristiky v kontextu jiných položek z této části maturitního testu nízká či vysoká, což usnadňuje její interpretaci.

Uvedené hodnoty mohou tvůrci testu využít k expertnímu posouzení, zda položka svou jazykovou úrovní odpovídá předpokládanému účelu testu, a případně k detailním úpravám některých parametrů (např. ke zkrácení textu nebo k nahrazení dlouhých slov kratšími synonymy, pokud se text položky na základě uvedených charakteristik jeví jako příliš složitý). Zatímco základní vlastnosti textu detailně charakterizují jednotlivé prvky jako slova nebo věty, indexy čitelnosti charakterizují celkovou míru složitosti textu pro čtenáře.

Item Wording Text Characteristi

Percentile: 18.8
Range: 665; 4139
Median: 1902.5

Item feature		Item passage	Item question	Correct option	All incorrect
Number of characters	999	815	73	30	137
Number of tokens	211	175	8	6	27
Word length standard deviation (characters)	1.8	1.8	1.5	0.6	1.7
Average word length (characters)	4.7	4.7	4.6	5	5.1
Longest word length (characters)	12	12	7	8	9
Text readability - FOG index	13.6	11.8	2.8	10	11.6
Text readability - Dale-Chall index	6.8	6.9	0.3	7	1.2
Text readability - Traenkle-Bailer index	-459.6	-398	-191.7	-287.4	-453.3
Text readability - SMOG index	11.2	10.6	3.1	8.8	8.8

Hover on the numbers to see the overal range and median for a given feature, including the percentile rank of the value relative to the items from the training dataset. Please see the article below for more details and explanations of each item feature. Exluded words (i.e. "stopwords"): I, a, about, an, are, as, at, be, by, com, for, from, how, in, is, it, of, on, or, that, the, this, to, was, what, when, where, who, will, with, www.

Obrázek 3: Výstup analýzy textových vlastností položek v modulu *EduTest Text Analysis*

Na obrázku 3 je zobrazena část výstupu z modulu, konkrétně přehled základních textových vlastností pro položku 25 z jarního termínu 2017. V této části jsou analyzovány základní lexikální vlastnosti textu (počet znaků, počet tokenů[2], průměrná délka slova atd.) a hodnoty indexů čitelnosti textu (Gunningův Fog index, index Dalea a Challové, index Tränkla a Bailera a index SMOG). Hodnoty jsou zobrazeny pro celý text položky (all text), ale také pro jednotlivé části položky, tedy výchozí text (item passage), otázku (item question), správnou možnost (correct option) a nesprávné možnosti neboli distraktory (all incorrect). Po najetí myší na jednotlivé hodnoty v tabulce lze získat podrobnější informace o percentilu, rozpětí a mediánu dané vlastnosti. Text zmíněné položky má celkem 999 znaků, z toho výchozí text 815 znaků, otázka 73 znaků, správná odpověď 30 znaků a všechny nesprávné odpovědi dohromady 137 znaků.

[2]V modulu *EduTest Text Analysis* jsou jako tokeny použita slova.

Celková délka textu odpovídá 19. percentilu, tedy v testech zadávaných v letech 2016–2023 mělo stejnou nebo nižší délku 19 % položek, jedná se tedy spíše o kratší položku. Pokud se podíváme na další charakteristiky, tak vidíme, že například průměrná délka slova ve znacích (average word length (characters)) je 4,7, což odpovídá pouze 10. percentilu, tedy v textu položky jsou průměrně kratší slova než v jiných položkách. Index čitelnosti Dalea a Challové má hodnotu 6,8, což také naznačuje nižší obtížnost textu (24. percentil). Při interpretaci indexů čitelnosti je třeba vzít v úvahu, že index Tränkla a Bailera je vyjádřen na převrácené škále než ostatní indexy, tj. vyšší hodnoty (ležící blíže k nule) znamenají lepší čitelnost. Nízké hodnoty percentilů indexu Tränkla a Bailera tedy naznačují, že položka klade ve srovnání s ostatními dostupnými položkami vyšší nároky na čitelnost.

3.4 Měření podobnosti textu: zkoumání lexikálních a sémantických vztahů

Kromě vlastností, které jsou podstatné pro srozumitelnost nebo čitelnost textu jako takového, může mít na obtížnost testových položek pro hodnocení porozumění textu specifický vliv také podobnost nebo odlišnost různých částí testové položky (např. výchozího textu a správné odpovědi, výchozího textu a distraktorů nebo správné odpovědi a distraktorů).

V našem výzkumu jsme rozlišili dva základní typy měření podobnosti textů: lexikální a sémantickou. Lexikální podobnost se zaměřuje na sdílenou slovní zásobu v různých částech textu (Tan et al., 2005). Většinou se jedná o podobnost textů především z hlediska skutečně použitých slov, která se zaměřuje na přítomnost stejných slov a slovních spojení ve srovnávaných textech. Nebere ale v úvahu kontext a význam slov. Sémantická podobnost naopak hlouběji proniká do porozumění významům a kontextům slov použitých v textech. Sémantická podobnost pracuje s myšlenkou, že dva texty, jejichž významy jsou si blízké, mohou mít vysokou podobnost i přesto, že nemají významný počet společných slov. Pro hodnocení sémantické podobnosti jsou k dispozici různé jazykové modely, včetně modelů založených na učení bez učitele. Tyto modely pracují s vektorovou reprezentací slov, což je způsob, jakým můžeme slova vyjádřit ve formě číselných vektorů. Tento přístup využívá velké množství textových dat, a umožňuje tak zachytit sémantický význam slov na základě kontextu, ve kterém se vyskytují.

Modul pro analýzu textů vyhodnocuje podobnost textů pomocí tří metrik:

1. **Kosinová podobnost (cosine similarity)**: Tento ukazatel měří lexikální podobnost mezi dvěma tokeny (v případě tohoto modulu jsou jako tokeny použita slova, ale tokeny mohou být i znaky, spojení n prvků – například slov, tzv. n-gramy, nebo celé věty), reprezentovanými pomocí numerických vektorů, založenou na kosinu úhlu, který mezi sebou svírají. Nižší hodnota (blížící se k 0) naznačuje větší rozdíl mezi vektory, tedy menší počet společných slov (Deza & Deza, 2016; Gunawan et al., 2018). Lze předpokládat, že např. nižší kosinová podobnost mezi výchozím textem a zněním správné odpovědi bude zvyšovat obtížnost položky, protože jejich souvislost nebude na první pohled zřejmá a žák bude muset oba texty číst důkladněji (Brizuela & Montero-Rojas, 2014, s. 6). Kosinová podobnost může být použita i jako měřítko sémantické podobnosti (Hvitfeldt & Silge, 2021), avšak v interaktivní aplikaci ji používáme pro měření lexikální podobnosti.

2. **Euklidovská vzdálenost (Euclidean distance)**: Tento ukazatel lexikální podobnosti měří doslovnou „vzdálenost“ mezi dvěma tokeny, tedy přímou „linii“ mezi těmito dvěma body, na rozdíl od kosinové podobnosti, která se více zaměřuje na úhel mezi vektory. Výsledná hodnota tedy poskytuje přímý odhad „vzdálenosti“ nebo rozdílu mezi dvěma texty na základě jejich reprezentace v n-rozměrném prostoru. Je to užitečný ukazatel pro měření odchylek nebo rozdílů mezi texty v případě, že jsou relevantní specifické rozdíly v jednotlivých textech. Vyšší euklidovská vzdálenost mezi výchozím textem a zněním

správné odpovědi může znamenat, že pro žáky bude obtížnější pochopit souvislost, a zvolit tak správnou odpověď. Na druhou stranu, vyšší euklidovská vzdálenost mezi výchozím textem a distraktory a mezi distraktory a správnou odpovědí může žákům napovědět která odpověď je správná, a vést tak k nižší obtížnosti položky (Deza & Deza, 2016, s. 103).

3. **Podobnost slovních vektorů (word vectors similarity)**: Tento ukazatel měří sémantickou podobnost mezi slovy nebo texty. V modulu je použit model Word2Vec (Mikolov et al., 2013), který má potenciál efektivně zachytit významovou podobnost ve stručných textech a poskytnout přesné výsledky i při omezené délce textu (tedy například v případě nabízených možností odpovědi). Model Word2Vec je zároveň efektivnější z hlediska výpočetních nároků a snadnosti implementace. Každé slovo je v tomto modelu převedeno na bod v prostoru s mnoha rozměry, což umožňuje studovat vzájemné vztahy mezi slovy na základě jejich polohy v tomto prostoru. Tento přístup umožňuje porozumět skryté sémantické struktuře textu, a tím přispět k odhadu jeho obtížnosti. Lze očekávat, že nižší významová podobnost mezi některými částmi položky může vést, podobně jako v případě kosinové podobnosti, k vyšší obtížnosti položky (Hvitfeldt & Silge, 2021; Mikolov et al., 2013).

V rámci interaktivního modulu je k dispozici výpočet několika metrik podobnosti či vzdálenosti mezi různými částmi testové položky vložené uživatelem (obrázek 4). Kromě kosinové podobnosti, euklidovské vzdálenosti a podobnosti slovních vektorů modul pro každou dvojici srovnávaných částí položky počítá také podíl společných slov.

Similarity Measures Between Item Wording Parts

Characteristics	Correct vs. incorrect options	Item passage vs. incorrect options	Item passage vs. correct option	Question vs. incorrect options	Question vs. correct option	Question vs. item passage
Common words$_1$	0.6	0.3	0.1	0	0	0.69
Common words$_2$	0.27	0.88	0.8	0	0	0.04
Cosine Similarity	0.47	Percentile: 99.4 Range: 0; 0.9 Median: 0.7	0.29	0.07	0	0.17
Euclidean Distance	5.92		24.35	7.81	5.1	24.56
Word2vec Similarity	0.56	0.92	0.54	0.71	0.37	0.75

Hover on the numbers to see the overal range and median for a given feature, including the percentile rank of the value relative to the items from the training dataset. Common words$_1$: proportion of common words from a text of part A found in a text of part B of the item wording; common words$_2$: proportion of common words from a text of part B found in a text of part A of the item wording.

Obrázek 4: Výstup analýzy podobnosti jednotlivých částí znění položky v modulu *EduTest Text Analysis*

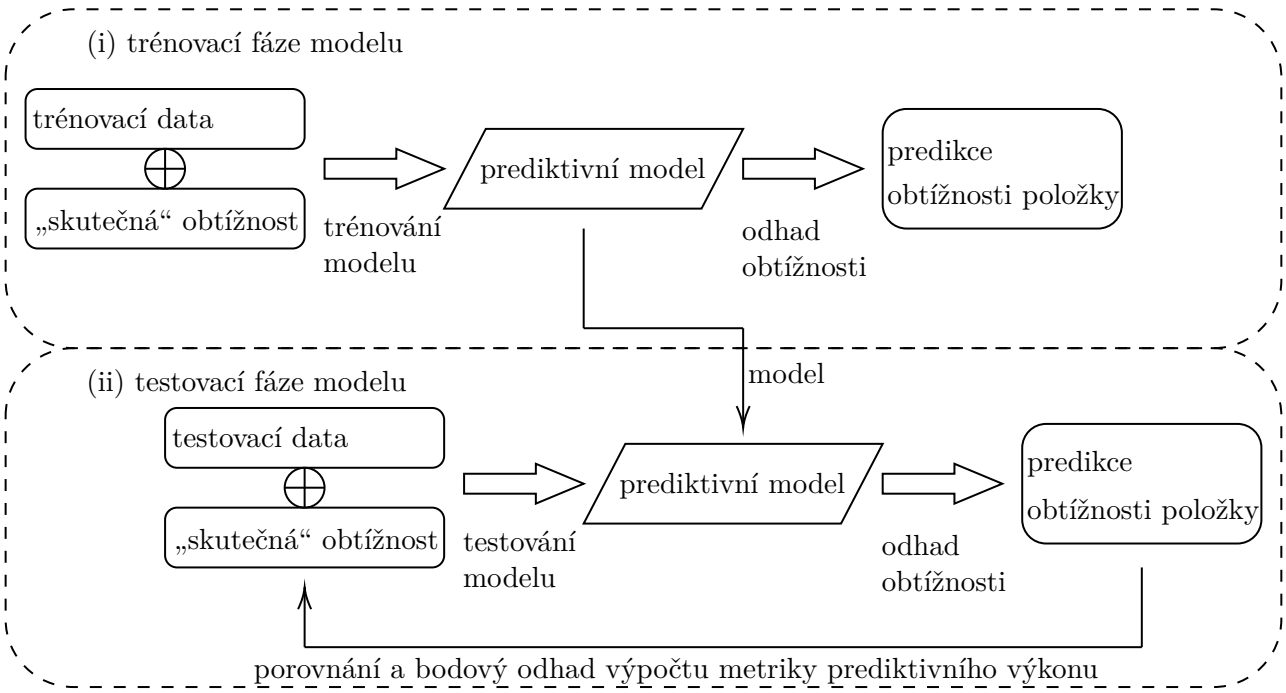
V této konkrétní položce (položka 25 z jarního termínu 2017) například vidíme, že znění výchozího textu má poměrně vysokou lexikální podobnost (kosinová podobnost 0,54, 85. percentil; euklidovská vzdálenost 22,05, 24. percentil) se zněním nesprávných odpovědí a sémantická podobnost mezi těmito dvěma částmi položky dosahuje dokonce jedné z nejvyšších hodnot ve srovnání s ostatními položkami z databáze (podobnost slovních vektorů 0,92, 99. percentil). V textu nesprávných odpovědí se také objevuje poměrně vysoký podíl slov z výchozího textu (0,3, 99. percentil). Oproti tomu doslovná i sémantická podobnost mezi výchozím textem a správnou odpovědí je v kontextu ostatních položek v databázi průměrná až nižší (kosinová podobnost 0,29, 59. percentil; podobnost slovních vektorů 0,54, 34. percentil). Tím klade tato položka nároky na pozorné čtení textu, protože vysoká podobnost výchozího textu se zněním distraktorů může při nepozorném čtení žáka zmást a vést k volbě nesprávné odpovědi.

4 Predikce obtížnosti položek pomocí algoritmů strojového učení

Výše představené textové vlastnosti položek lze nejen vyhodnocovat jednotlivě, jak demonstrujeme v sekci 3, ale také je lze využít k automatizované predikci obtížnosti položek pomocí algoritmů strojového učení, kterou představíme v této sekci. Vychází se přitom z tzv. trénovacích dat, pro která je kromě textových vlastností položek k dispozici také odhad jejich obtížnosti na základě odpovědí žáků, kteří již na položky z trénovacího datasetu odpovídali. V naší práci využíváme odhad obtížnosti položek na základě Raschova (tedy jednoparametrického logistického) modelu, používáme metodu podmíněného maximálního věrohodnostního odhadu (Martinková & Hladká, 2023, s. 165). Takto získaný odhad obtížnosti položek se snažíme co nejlépe predikovat textovými vlastnostmi položek.

Pro predikci obtížnosti položek v interaktivním modulu byla využita regresní metoda označovaná jako elastická síť (anglicky *elastic net regression*, zkráceně *elnet*), která je zvláště užitečná při práci s datovými soubory obsahujícími mnoho proměnných, protože efektivně hledá „rovnováhu“ mezi výběrem proměnných, v našem případě textových charakteristik položek, které jsou důležité pro dobrý prediktivní výkon modelu. Elastická síť přidává k obvyklé lineární regresní rovnici dvě regularizační, resp. penalizační podmínky, což algoritmu umožňuje penalizovat ty modely, které obsahují proměnné, tedy textové charakteristiky položek, jež jsou pro prediktivní výkon modelu irelevantní nebo méně důležité. Takový přístup pomáhá zabránit přeučení (*overfitting*), tedy situaci, kdy model sice velmi dobře predikuje na trénovacím datasetu, ale obvykle selhává v predikci na jiných datech. Elastická síť je regresní technika, která kombinuje vlastnosti jiných dvou regularizačních regresních metod, a sice LASSO (Least Absolute Shrinkage and Selection Operator) regrese a hřebenové regrese (známé též jako *ridge* regrese). Díky kombinaci penalizačních podmínek z obou těchto metod je elastická síť v predikci obvykle minimálně tak efektivní jako dílčí dva modely, tedy LASSO regrese a hřebenová regrese.

V první fázi byl model elastické sítě natrénován na datasetu obsahujícím všechny položky zadané v jarních termínech 2016–2023. Při trénování modelu bylo vždy vybráno jiných devatenáct dvacetin datasetu, které sloužily jako trénovací dataset, a na zbylé dvacetině, tedy testovacím datasetu, byl model testován. Tento proces, nazývaný *křížová validace*, byl dvacetkrát opakován, což umožnilo získat robustnější odhady prediktivního výkonu, viz obrázek 5.



Obrázek 5: Diagram predikce obtížnosti položky pomocí metody strojového učení. (i) model byl trénován na základě trénovacího datasetu, (ii) model je použit pro slovní zadání nové položky, vložené do modulu. Predikovaná hodnota obtížnosti položky je pak porovnávána se „skutečnou“ hodnotou odhadnutou pomocí Raschova modelu. „Skutečné“ hodnoty jsou uváděny v uvozovkách, protože jde o hodnoty odhadnuté modelem, nikoliv známé; vysvětlení viz poznámku pod čarou. Dle Štěpánek et al. (2023), upraveno.

Současně se v rámci každé iterace křížové validace model opakovaně trénoval pro jiné kombinace parametrů penalizačních podmínek; výsledkem tedy byla velká řada potenciálních modelů s různými hodnotami penalizačních podmínek, trénovaných na různých, ale stejně velkých trénovacích datasetech. Jako finální model pro predikci byl pak automaticky vybrán ten, který minimalizuje predikční chybu, zde počítanou jako střední čtvercovou chybu, resp. její odmocninu, běžně označovanou též (root) mean square error ((R)MSE). Střední čtvercovou chybu, resp. její odmocninu, si lze rámcově představit jako rozptyl, resp. směrodatnou odchylku, rozdílů predikovaných a „skutečných“[3] hodnot obtížnosti položek.

V interaktivním modulu je obtížnost položek predikována pomocí finálního prediktivního modelu, který ze všech potenciálních modelů předpovídal obtížnost položek nejpřesněji. Parametry tohoto modelu jsou uvedeny v tabulce 1. Výsledný optimální model predikuje vyšší obtížnost položky, která má vyšší počet znaků, vyšší směrodatnou odchylku délky slov, vyšší počet tokenů v distraktorech, vyšší index Dalea a Challové, vyšší Fog index, více společných slov ve výchozím textu a distraktorech a vyšší podobnost správné odpovědi a distraktorů měřenou modelem Word2Vec.[4]

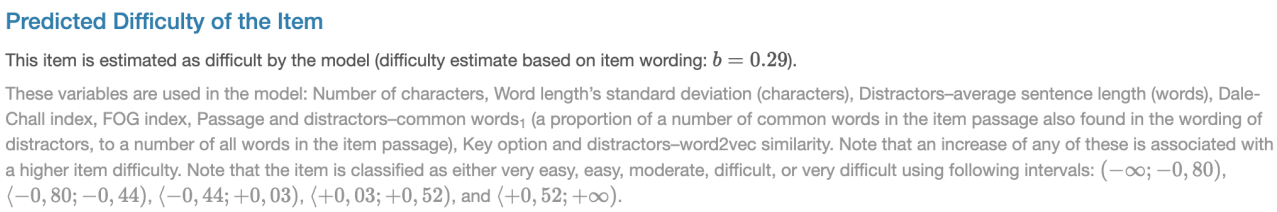
[3] „Skutečné“ hodnoty záměrně uvádíme v uvozovkách, protože skutečné hodnoty obtížnosti testových položek nejsou známy. Pracujeme s hodnotami odhadnutými Raschovým modelem, které mohou do predikce vnášet další stupeň nejistoty, ačkoliv je považujeme za nejlepší odhad obtížnosti.

[4] V rámci interaktivního modulu jsou drobné odchylky v implementaci modelu oproti původní studii. Aby bylo možné efektivně použít odhad podobnosti v modulu, byly z původně použitého modelu Word2Vec odstraněny n-gramy, které nejsou v našem případě relevantní, jelikož je v současné době nevyužíváme. Dále byly odebrány tokeny obsahující čísla, speciální znaky a ty, které se skládaly pouze z velkých písmen. Tento krok byl proveden s cílem eliminovat nežádoucí šum a přizpůsobit model tak, aby byl více zaměřený na čistě textová data. Zjednodušený model, který nyní obsahuje pouze 750 000 tokenů z původních přibližně 3 milionů, umožňuje zrychlit výpočty při zachování přesnosti a snížené zátěže serveru. Rozdíly v predikovaných hodnotách obtížnosti byly na vybraných položkách minimální (v řádu setin).

Tabulka 1: Koeficienty modelu elastické sítě, který minimalizuje RMSE s $\hat{\lambda}_{\text{LASSO}} \approx 1$ a $\hat{\lambda}_{\text{ridge}} \approx 0$. Dle Štěpánek et al. (2023).

textová charakteristika položky	koeficient
(intercept)	-3,808
počet znaků	0,002
směrodatná odchylka délky slov (počtu znaků)	0,809
počet tokenů v distraktorech	0,002
index Dalea a Challové	0,004
Fog index	0,026
společná slova výchozího textu a distraktorů	0,630
Word2Vec podobnost správné odpovědi a distraktorů	0,023

Výsledek predikce pro novou položku, jejíž textové zadání bylo vloženo do modulu, je v interaktivním modulu zobrazen v dolní části výstupu (obrázek 6). Kromě predikce obtížnosti položky jako číselné hodnoty, obdobně jako je odhadována Raschovým modelem, je možné položku zatřídit i do kategorie podle obtížnosti. Konkrétně rozlišujeme pět kategorií obtížnosti položek, tj. položku můžeme co do obtížnosti vyhodnotit jako velmi lehkou (very easy), lehkou (easy), středně těžkou (moderate), obtížnou (difficult) a velmi obtížnou (very difficult). Kategorie obtížností jsou navrženy tak, aby do každé kategorie patřila zhruba pětina všech položek z trénovacího datasetu. Hranice mezi intervaly, na které je spojitá škála obtížnosti rozdělena, tedy odpovídají kvintilům rozsahu obtížnosti.[5]



Obrázek 6: Výsledek odhadu obtížnosti v modulu *EduTest Text Analysis*

Ukázková položka (položka 25 z jarního termínu 2017) má modelem predikovanou obtížnost $b = 0{,}29$, což odpovídá intervalu $\langle +0{,}03; +0{,}52)$, tedy tato položka je v kontextu položek z trénovacího datasetu vyhodnocena jako obtížná (obrázek 6).

5 Praktická ukázka

Pro ilustraci vztahu mezi vlastnostmi textu a obtížností testových položek si ukážeme analýzu dvou konkrétních položek. Položka číslo 28 z jarního termínu 2017 (obrázek 7), se zabývá otázkou o tajné zprávě nalezené spolu s kostrou poštovního holuba. V kontextu položek zadávaných v letech 2016–2023 má tato položka vysokou obtížnost ($b = 1{,}63$ na základě empirických žákovských odpovědí).

[5]Konkrétní intervaly jsou následující: $(-\infty; 0{,}80)$, $\langle -0{,}80; -0{,}44)$, $\langle -0{,}44; +0{,}03)$, $\langle +0{,}03; +0{,}52)$ a $\langle +0{,}52; +\infty)$.

World War II Message Found

Last month, David Martin, a car mechanic from southern England, found the skeleton of a homing pigeon1 with a message in the chimney while he was renovating his house. During World War II (WWII), pigeons like this one were taken to the Nazi-occupied territories in France and sent back to Britain with messages from the British Army. Historians believe that this pigeon began its journey in France in June 1944 and never arrived to deliver its message. The pigeon either got lost in the bad English weather or was tired after its trip across the Channel and ended up falling down the chimney. The message that was found by David Martin next to the pigeon was certainly secret because it was written in a strange code. Now historians are trying to understand the message and find out whether it could possibly have changed history.

(www.dailymail.co.uk, upraveno)

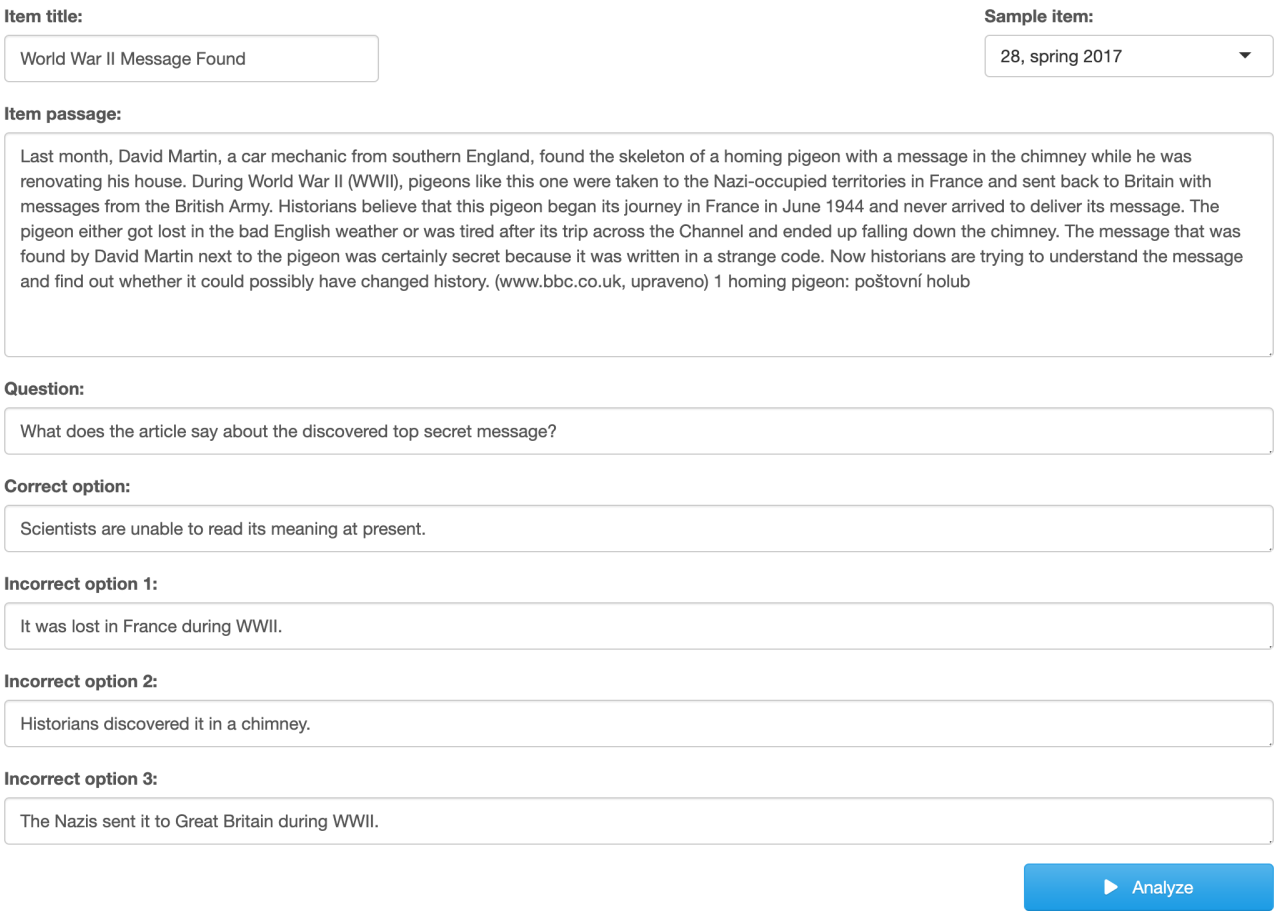
[1]homing pigeon: poštovní holub

28 What does the article say about the discovered top secret message?

A) It was lost in France during WWII.
B) Historians discovered it in a chimney.
C) The Nazis sent it to Great Britain during WWII.
D) Scientists are unable to read its meaning at present.

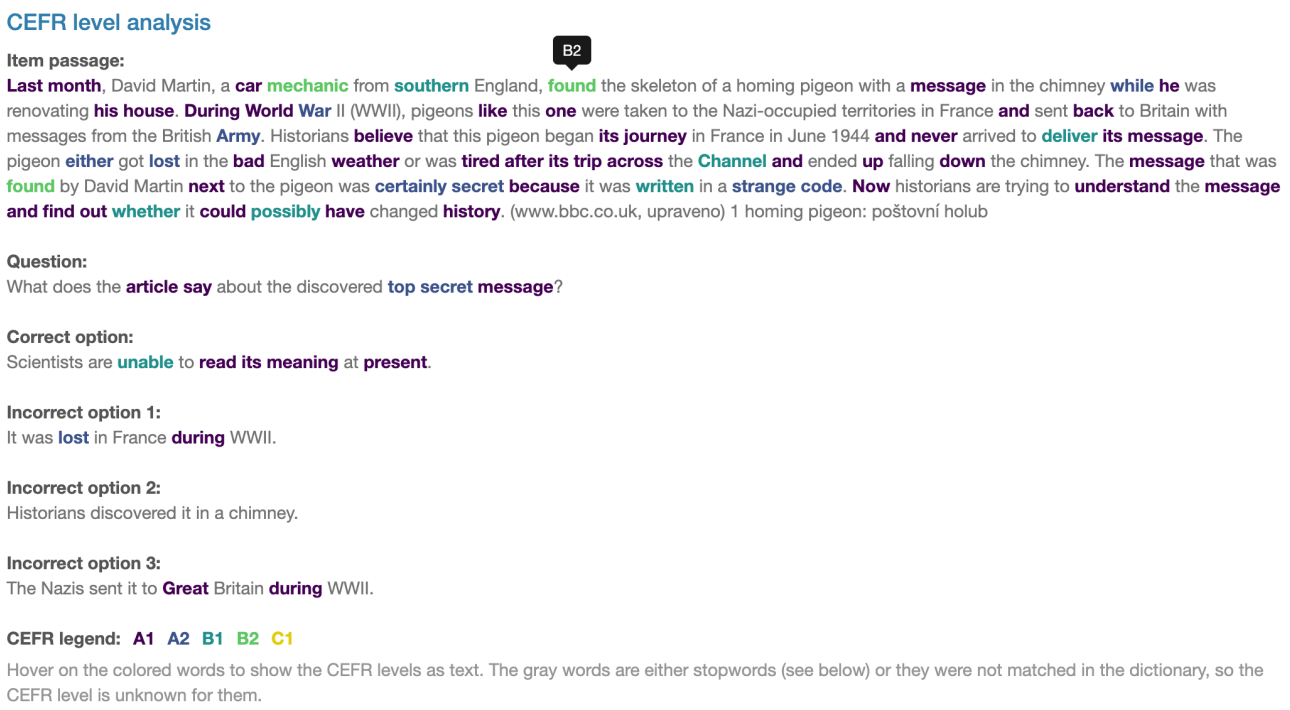
Obrázek 7: Ukázka položky 28 z jarního termínu 2017 s vyšší obtížností ($b = 1{,}63$ na základě empirických žákovských odpovědí), správná odpověď D

Jednotlivé části znění položky se nakopírují do modulu *EduTest Text Analysis* (obrázek 8) a pomocí tlačítka **Analyze** se spustí analýza.



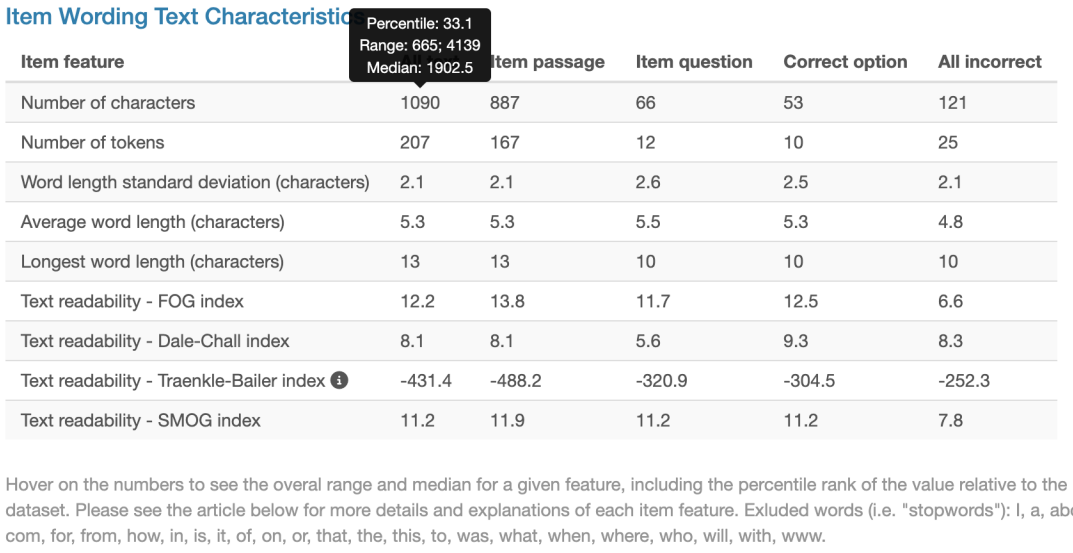
Obrázek 8: Vložení znění položky 28 z jarního termínu 2017

CEFR analýza označila úrovně jednotlivých slov (obrázek 9). Většina slov je úrovně A1, A2 nebo B1, což je v souladu s předpokládanou úrovní obtížnost testu.



Obrázek 9: CEFR analýza položky 28 z jarního termínu 2017

Tabulka s textovými charakteristikami (obrázek 10) naznačuje, že se jedná o poměrně krátkou položku (1090 znaků, 33. percentil; 207 tokenů, 37. percentil). Položka má také relativně nízkou variabilitu délky slov – směrodatná odchylka délky slov je 2,1 znaku (27. percentil). Zároveň ale obsahuje delší slova: průměrná délka slova je 5,3 znaku (92. percentil) a nejdelší slovo má 13 znaků (83. percentil). Všechny indexy čitelnosti jsou vysoké: indexy Dalea a Challové (8,1) a Fog (12,2) odpovídají 84. percentilu, index SMOG (11,2) 71. percentilu a index Tränkla a Bailera, který je vyjádřen na obrácené škále, odpovídá 8. percentilu (tj. pouze 8 % položek z databáze má stejnou nebo horší čitelnost).



Item feature		Item passage	Item question	Correct option	All incorrect
Number of characters	1090	887	66	53	121
Number of tokens	207	167	12	10	25
Word length standard deviation (characters)	2.1	2.1	2.6	2.5	2.1
Average word length (characters)	5.3	5.3	5.5	5.3	4.8
Longest word length (characters)	13	13	10	10	10
Text readability - FOG index	12.2	13.8	11.7	12.5	6.6
Text readability - Dale-Chall index	8.1	8.1	5.6	9.3	8.3
Text readability - Traenkle-Bailer index	-431.4	-488.2	-320.9	-304.5	-252.3
Text readability - SMOG index	11.2	11.9	11.2	11.2	7.8

Hover on the numbers to see the overal range and median for a given feature, including the percentile rank of the value relative to the items from the training dataset. Please see the article below for more details and explanations of each item feature. Exluded words (i.e. "stopwords"): I, a, about, an, are, as, at, be, by, com, for, from, how, in, is, it, of, on, or, that, the, this, to, was, what, when, where, who, will, with, www.

Obrázek 10: Textové charakteristiky položky 28 z jarního termínu 2017

Tabulka podobností mezi jednotlivými částmi (obrázek 11) poskytuje další informace o znění

položky. Tato položka vykazuje středně silnou kosinovou podobnost (0,29, 59. percentil) a podobnost slovních vektorů (0,66, 59. percentil) mezi výchozím textem a správnou odpovědí a velmi nízkou euklidovskou vzdálenost (19,18, 9. percentil) mezi těmito dvěma částmi textu. Zároveň je zde patrná vysoká kosinová podobnost (0,5, 79. percentil) a podobnost slovních vektorů (0,9, 96. percentil) mezi výchozím textem a distraktory. Euklidovská vzdálenost je i v tomto případě velmi nízká (15,6, 5. percentil). Tyto hodnoty naznačují vysokou podobnost mezi výchozím textem a distraktory, zatímco vztah mezi výchozím textem a správnou odpovědí je méně jasný. To potvrzují i podíly společných slov – podíl slov z výchozího textu, které se nacházejí ve znění distraktorů, je 0,21 (84. percentil), zatímco ve správné odpovědi se nachází jen 11 % slov z výchozího textu (75. percentil). Tyto faktory mohou způsobovat vyšší obtížnost položky.

Similarity Measures Between Item Wording Parts

Characteristics	Correct vs. incorrect options	Item passage vs. incorrect options	Item passage vs. correct option	Question vs. incorrect options	Question vs. correct option	Question vs. item passage
Common words$_1$	0.22	0.21	0.11	0.1	0	0.2
Common words$_2$	0.18	0.73	0.33	0.05	0	0.04
Cosine Similarity	0.21	Percentile: 95.6 Range: 0; 0.9 Median: 0.7	0.29	0.05	0	0.1
Euclidean Distance	6.4		19.18	6.93	4.36	19.77
Word2vec Similarity	0.58	0.9	0.66	0.72	0.6	0.8

Hover on the numbers to see the overal range and median for a given feature, including the percentile rank of the value relative to the items from the training dataset. Common words$_1$: proportion of common words from a text of part A found in a text of part B of the item wording; common words$_2$: proportion of common words from a text of part B found in a text of part A of the item wording.

Obrázek 11: Výstup analýzy podobnosti pro položku 28 z jarního termínu 2017

Odhad obtížnosti položky pomocí regresního modelu je 0,62 (obrázek 12), což spadá do nejvyšší kategorie obtížnosti. Rozdíl mezi „skutečnou“ obtížností položky získanou z žákovských odpovědí a predikcí její obtížnosti na základě textu není příliš velký, položka byla oběma přístupy klasifikována jako velmi obtížná.

Predicted Difficulty of the Item

This item is estimated as very difficult by the model (difficulty estimate based on item wording: $b = 0.62$).

These variables are used in the model: Number of characters, Word length's standard deviation (characters), Distractors–average sentence length (words), Dale-Chall index, FOG index, Passage and distractors–common words$_1$ (a proportion of a number of common words in the item passage also found in the wording of distractors, to a number of all words in the item passage), Key option and distractors–word2vec similarity. Note that an increase of any of these is associated with a higher item difficulty. Note that the item is classified as either very easy, easy, moderate, difficult, or very difficult using following intervals: $(-\infty; -0,80)$, $\langle -0,80; -0,44)$, $\langle -0,44; +0,03)$, $\langle +0,03; +0,52)$, and $\langle +0,52; +\infty)$.

Obrázek 12: Predikce obtížnosti pro položku 28 z jarního termínu 2017

Jako druhý příklad analýzy obtížnosti položky na základě textu jsme vybrali položku číslo 25 z jarního termínu 2022 (obrázek 13). Tato položka se týká incidentu s tygrem, který přišel k lidem pro pomoc. V rámci položek zadávaných v maturitních testech z anglického jazyka v letech 2016–2023 se jedná o položku s nízkou obtížností (na základě empirických žákovských odpovědí byla její odhadnutá obtížnost $b = -2,48$).

The Tiger which Asked for Help

One day, a tiger appeared at Ivan Pavlov's house in Solontsovyy, a Russian village. 'It looked very weak and I thought it had been hurt, maybe a car hit it as you don't expect to see a tiger so close to the village. I called emergency services,' says Pavlov. They arrived quickly and moved the tiger to a veterinary clinic. At first they thought the tiger had been hurt by hunters, but when they looked closer at its body, they found no injuries. There were no broken bones, either, which meant that Pavlov was wrong. But when vets opened the tiger's mouth, they were surprised: its upper teeth had fallen out. The vets believe the animal came to people for help because without teeth it could no longer hunt. Now they have to decide if they should move the tiger to a zoo. They are worried other tigers might hurt it or even worse – kill it, which isn't unusual behaviour.

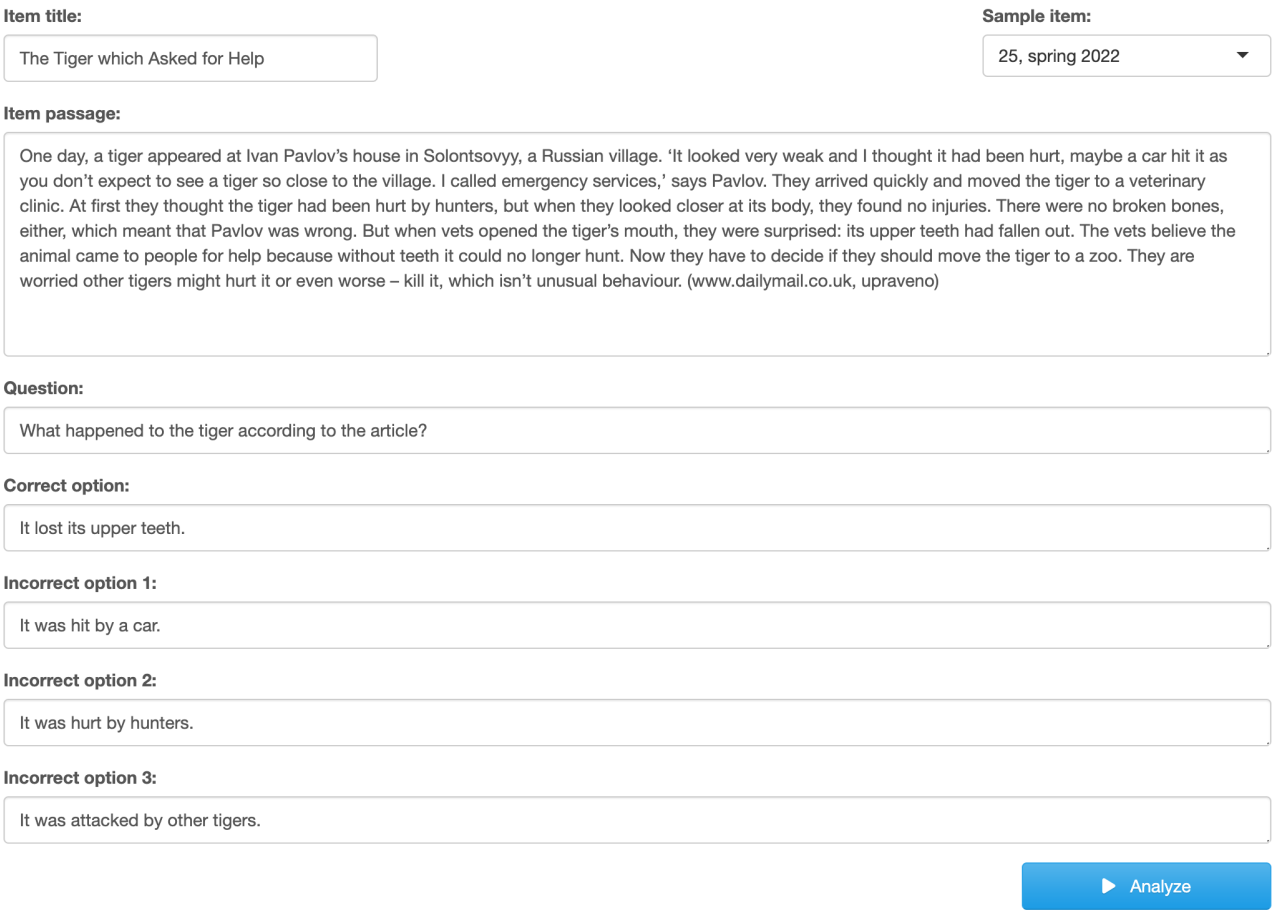
(www.dailymail.co.uk, upraveno)

25 What happened to the tiger according to the article?

A) It was hit by a car.
B) It lost its upper teeth.
C) It was hurt by hunters.
D) It was attacked by other tigers.

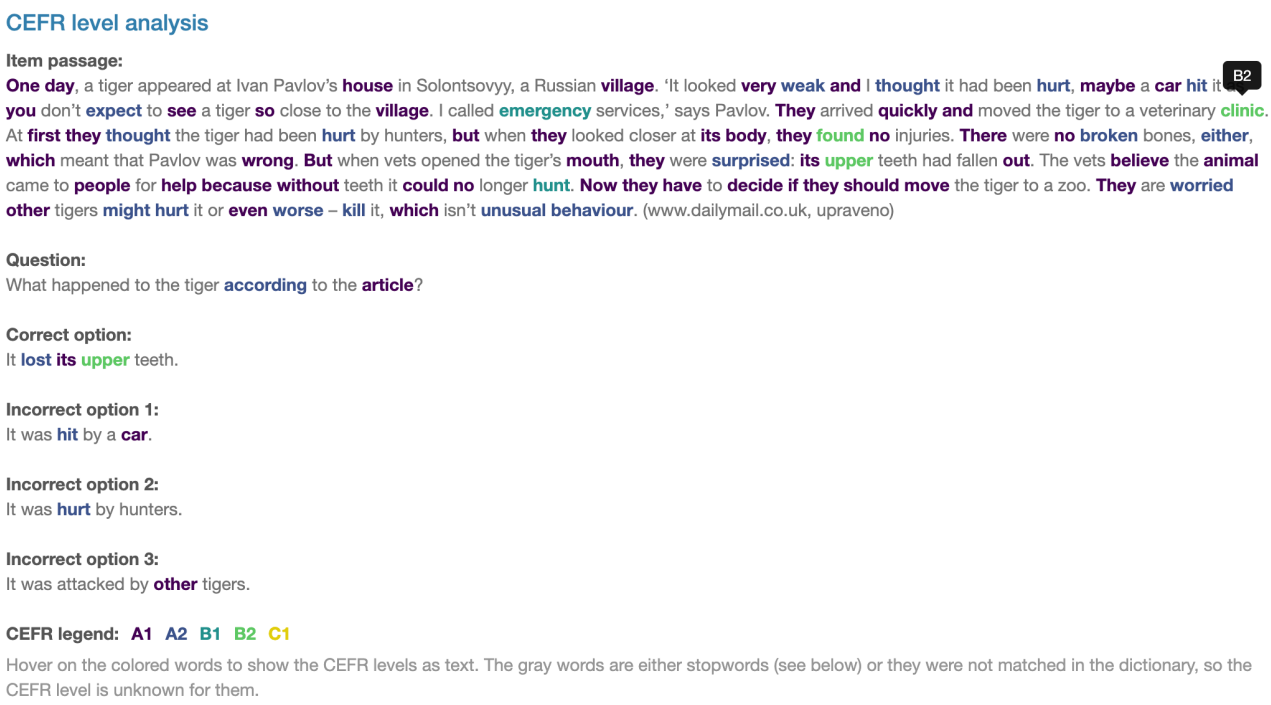
Obrázek 13: Ukázka položky 25 z jarního termínu 2022 s nízkou obtížností ($b = -2{,}48$ na základě empirických žákovských odpovědí), správná odpověď B

Po vložení jednotlivých částí znění položky do modulu *EduTest Text Analysis* (obrázek 14) se pomocí tlačítka **Analyze** spustí textová analýza.



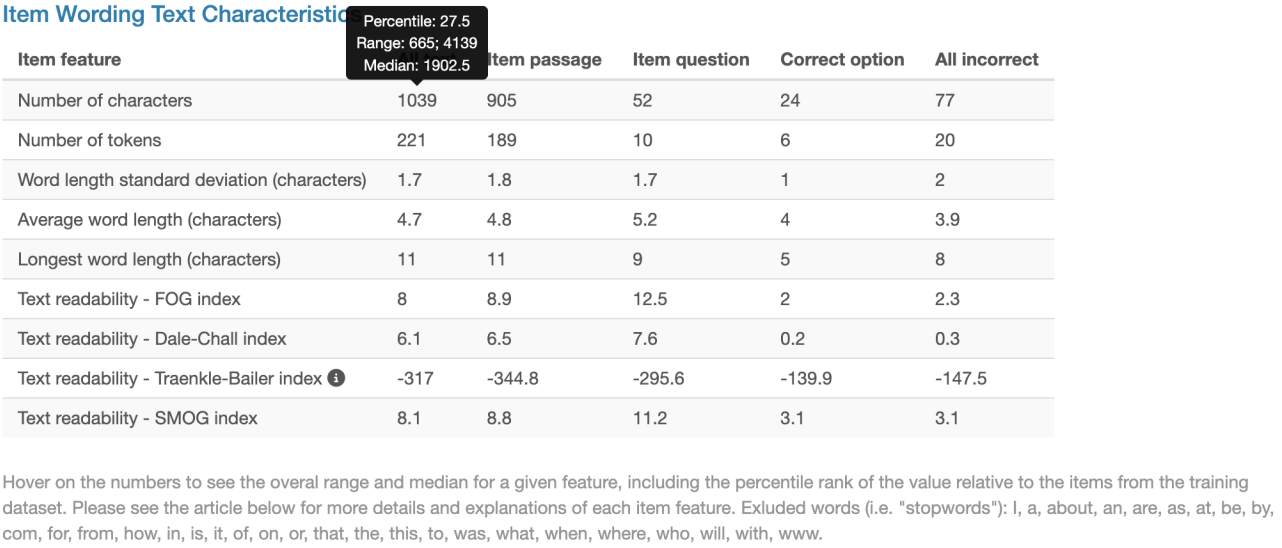
Obrázek 14: Vložení položky 25 z jarního termínu 2022

CEFR analýza označila úrovně jednotlivých slov (obrázek 15). Většina slov je úrovně A1, A2 nebo B1.



Obrázek 15: CEFR analýza položky 25 z jarního termínu 2022

Tabulka s textovými charakteristikami (obrázek 16) naznačuje, že by se mohlo jednat o poměrně snadnou položku. Počet znaků je celkem 1039, což odpovídá 28. percentilu. Text je tedy spíše kratší. Variabilita délky slov v textu je velmi nízká (směrodatná odchylka je 1,7 znaku, 1. percentil), velmi nízká je také průměrná délka slova (4,7 znaku, 10. percentil). Rovněž indexy čitelnosti naznačují nízkou složitost textu (index Dalea a Challové je 6,1, což odpovídá 4. percentilu, hodnota indexu Fog je 8, což odpovídá 5. percentilu). Všechny tyto charakteristiky naznačují, že čitelnost textu bude zřejmě velmi dobrá.



Item feature		Item passage	Item question	Correct option	All incorrect
Number of characters	1039	905	52	24	77
Number of tokens	221	189	10	6	20
Word length standard deviation (characters)	1.7	1.8	1.7	1	2
Average word length (characters)	4.7	4.8	5.2	4	3.9
Longest word length (characters)	11	11	9	5	8
Text readability - FOG index	8	8.9	12.5	2	2.3
Text readability - Dale-Chall index	6.1	6.5	7.6	0.2	0.3
Text readability - Traenkle-Bailer index	-317	-344.8	-295.6	-139.9	-147.5
Text readability - SMOG index	8.1	8.8	11.2	3.1	3.1

Obrázek 16: Textové charakteristiky položky 25 z jarního termínu 2022

Tabulka podobností mezi jednotlivými částmi (obrázek 17) poskytuje další informace o znění položky. Zde vidíme vyšší kosinovou podobnost (0,37, 76. percentil), středně vysokou podobnost

slovních vektorů (0,66, 59. percentil) a naopak nižší euklidovskou vzdálenost (25,7, 41. percentil) mezi výchozím textem a správnou odpovědí. Tyto hodnoty poukazují na jasnější spojení mezi textem a správnou odpovědí, což může přispívat k nižší obtížnosti položky. Podobné hodnoty ale pozorujeme i v případě vztahu výchozího textu s nesprávnými možnostmi, což naopak spíše zvyšuje obtížnost položky. Stejně tak je pro položku charakteristický poměrně vysoký podíl společných slov mezi výchozím textem a distraktory (0,16, 71. percentil) i mezi výchozím textem a správnou odpovědí (0,1, 69. percentil).

Similarity Measures Between Item Wording Parts

Characteristics	Correct vs. incorrect options	Item passage vs. incorrect options	Item passage vs. correct option	Question vs. incorrect options	Question vs. correct option	Question vs. item passage
Common words$_1$	0.33	0.16	0.1	0.12	0	0.38
Common words$_2$	0.38	0.94	0.83	0.06	0	0.07
Cosine Similarity	0.42	Percentile: 66.2 Range: 0; 0.9 Median: 0.7	0.37	0.05	0	0.23
Euclidean Distance	5.29		25.5	6.48	4	25.77
Word2vec Similarity	0.52	0.81	0.66	0.63	0.5	0.74

Hover on the numbers to see the overal range and median for a given feature, including the percentile rank of the value relative to the items from the training dataset. Common words$_1$: proportion of common words from a text of part A found in a text of part B of the item wording; common words$_2$: proportion of common words from a text of part B found in a text of part A of the item wording.

Obrázek 17: Výstup analýzy podobnosti pro položku 25 z jarního termínu 2022.

Odhad obtížnosti položky pomocí regresního modelu je 0,03 (obrázek 18), což je hodnota na hranici mezi kategoriemi středně těžká a obtížná. Rozdíl mezi „skutečnou“ obtížností položky získanou z žákovských odpovědí a predikcí obtížnosti položky na základě textu byl velký. Pro žáky byla položka velmi lehká, ale predikovaná obtížnost 0,03 řadí položku spíše mezi průměrně obtížné.

Vysvětlení je třeba hledat v parametrech zohledňovaných při predikci. Model pro predikci obtížnosti na základě znění položky zohledňuje více textových vlastností, ale nejvyšší váhu dává směrodatné odchylce délky slov a podílu společných slov z výchozího textu, které se nacházejí také ve znění distraktorů. Tento podíl je v dané položce vysoký (0,16, 71. percentil), což může zvyšovat nároky na pozornost čtenářů. V daném případě se však žáci nenechali zmást doslovnou podobností mezi distraktory a nesouvisejícími částmi výchozího textu a dokázali snadno identifikovat podstatu popsané události. Na správné místo výchozího textu je mohlo navést i slovní spojení „upper teeth“ vyskytující se v identickém znění jak ve výchozím textu, tak ve znění správné odpovědi. Zároveň je text celkově velmi dobře čitelný, což usnadňuje orientaci v textu a dohledávání informací o souvislostech, v nichž byla ve výchozím textu zmíněna slova z nabízených možností odpovědi, jako je „car“ nebo „hunters“. Čitelnost textu je v regresním modelu také zahrnuta, ale s mnohem menší vahou než variabilita délky slov a podíl společných slov. Pokud je snadná čitelnost podstatným faktorem pro porozumění textu položky, model to dostatečně nedocení. Predikční schopnost modelu navíc limitují další vlastnosti, které mohou ovlivňovat míru porozumění, ale analýza je nezohledňuje – například zajímavost textu nebo použité gramatické struktury.

Predicted Difficulty of the Item

This item is estimated as difficult by the model (difficulty estimate based on item wording: $b = 0.03$).

These variables are used in the model: Number of characters, Word length's standard deviation (characters), Distractors–average sentence length (words), Dale-Chall index, FOG index, Passage and distractors–common words$_1$ (a proportion of a number of common words in the item passage also found in the wording of distractors, to a number of all words in the item passage), Key option and distractors–word2vec similarity. Note that an increase of any of these is associated with a higher item difficulty. Note that the item is classified as either very easy, easy, moderate, difficult, or very difficult using following intervals: $(-\infty; -0,80)$, $\langle -0,80; -0,44)$, $\langle -0,44; +0,03)$, $\langle +0,03; +0,52)$, and $\langle +0,52; +\infty)$.

Obrázek 18: Predikce obtížnosti pro položku 25 z jarního termínu 2022

6 Diskuse a závěr

V tomto dokumentu jsme představili modul pro využití textové analýzy pro predikci obtížnosti položek. Modul v první fázi počítá jednotlivé textové charakteristiky, které jsou následně využity při odhadu obtížnosti s využitím algoritmů strojového učení. Odhad obtížnosti v modulu je založen na regresní metodě elastické sítě, která z velkého množství dostupných textových charakteristik efektivně vybírá ty, které jsou důležité pro dobrý prediktivní výkon modelu. Náš model je natrénovaný přímo na položkách testu z angličtiny zadaných v rámci maturitní zkoušky v předchozích letech (Štěpánek et al., 2023), je proto relevantní především pro položky zadávané v rámci tohoto testu v budoucnu.

Podobně jako modely natrénované na jiných datech v rámci zahraničních studií, i náš model předpovídá, že čím delší je text a čím více obsahuje složitých a neobvyklých slov, tím je položka obtížnější. Složitost a neobvyklost slov však není v modelu zohledňována přímo, ale zprostředkovaně přes index čitelnosti Dalea a Challové, do jehož výpočtu vstupuje podíl méně známých slov v textu, a Fog index, který zohledňuje poměr kratších a delších slov. Velmi silný vztah s obtížností položky má v našem modelu směrodatná odchylka délky slov, která jiným způsobem odkazuje k vyššímu podílu dlouhých, tj. složitých slov v textu položky. Vztah mezi indexy podobnosti a obtížností položky není zcela přímočarý a závisí na tom, které části položky jsou si vzájemně podobné. V našem případě do modelu vstupuje podíl společných slov výchozího textu a distraktorů a podobnost správné odpovědi a distraktorů, jejichž vyšší úroveň činí položku pro respondenta složitější. Silnější váhu má však podíl společných slov výchozího textu a distraktorů, což naznačuje klíčovou roli doslovné podobnosti v tomto typu testových položek.

Modul může sloužit jako podpůrný nástroj pro metodiky a předmětové koordinátory, kteří v rámci CZVV položky připravují a vytváří testy tak, aby jejich obtížnost byla meziročně co nejvíce srovnatelná. Ve stávající verzi modul pracuje pouze s položkami z anglického jazyka, pro který je nicméně k dispozici dostatek pretestových dat. V tomto ohledu má prezentovaný modul větší potenciál, pokud bude rozšířen na predikci obtížnosti úloh z německého, francouzského či dalších jazyků, pro které je získání pretestových dat požadované kvality obtížnější.

Náš výzkum je limitován nízkým počtem položek, na kterých mohl být regresní model trénován. Dalším omezením byla nízká variabilita obtížnosti položek (předepsaná obtížnost maturitního testu je B1, tedy poměrně nízká obtížnost v porovnání s průměrnou znalostí anglického jazyka v české populaci maturantů). To mohlo negativně ovlivnit vlastnosti modelu, v případě většího počtu položek nebo větší heterogenity jejich obtížností bychom mohli očekávat přesnější model.

K vylepšení prediktivního modelu by mohlo dále přispět zakomponování dalších charakteristik textu, včetně sémantických nebo obsahových. Z neformálních diskusí s tvůrci testů lze usuzovat, že obtížnost položek může být ovlivněna výskytem specifických jazykových jevů, které zatím v modelu nezohledňujeme (např. výskyt dvojího záporu činí položku obtížnější), resp. že obtížnost položky může také být ovlivněna zajímavostí a poutavostí tématu, kterého se text týká. V případě výpočtu složitějších textových charakteristik však v současné době narážíme na malý vzorek trénovacích dat (tj. nízký počet dosud zadaných položek daného typu), jak bylo naznačeno výše.

Velký potenciál by mohly mít rychle se rozvíjející jazykové modely, např. GloVe, nebo hluboce kontextualizované modely, jako jsou ELMo a BERT, které využívají složité neuronové sítě k pochopení kontextových významů slov (Devlin et al., 2018; Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018). Lepší reprezentaci krátkých textů může poskytnout také SBERT (Reimers & Gurevych, 2019). Další možné vylepšení by mohl poskytnout také model FastText, který je rozšířením modelu Word2Vec a je zajímavý svou schopností zpracovávat slova, která nejsou zahrnuta ve slovníku, pomocí rozkladu na menší slovní jednotky (Bojanowski et al.,

2016).

Dalším aspektem stávajícího výzkumu je skutečnost, že pro trénování modelu využíváme odhad obtížnosti položky na základě odpovědí respondentů. Tento odhad je sám o sobě do jisté míry nepřesný. V našem výzkumu se opíráme o vzorek všech maturantů, kteří konali zkoušku z anglického jazyka. V tomto kontextu je třeba poznamenat, že odhady obtížnosti položek na základě žákovských odpovědí jsou blízké skutečným obtížnostem položek, pouze pokud je k dispozici reprezentativní a dostatečně velký vzorek respondentů. Tak tomu bylo i v našem výzkumu, nicméně takový vzorek respondentů nemusí být k dispozici ve všech situacích. Např. pokud by byla místo kompletních ostrých dat použita data z pilotáže, spolehlivost dat použitých k trénování modelu by byla nižší.

Náš odhad obtížnosti je založen na Raschově IRT modelu a využívá metody podmíněné maximální věrohodnosti, která je výrazně méně náchylná k chybám odhadu způsobeným výraznými odlišnostmi mezi respondenty v jednotlivých zkušebních termínech. V budoucnosti by bylo vhodné zakomponovat např. metody vyrovnávání testů (test equating), díky kterým by mohlo být zachycení těchto odlišností ještě efektivnější.

Další možností do budoucna je zakomponovat do modelu pro odhad obtížnosti testových položek také expertní odhady tvůrců testů. V rámci interaktivního modulu by experti mohli vložit svůj odhad obtížnosti položky, který by byl následně využit pro korekci odhadu obtížnosti pomocí modelu.

Literatura

Alkhuzaey, A., & Tendeiro, J. N. (2020). A systematic review of data-driven approaches to item difficulty prediction. *Journal of Educational Measurement*, *57*(2), 263–280. https://doi.org/10.1111/jedm.12236

Beinborn, L., Zesch, T., & Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, *2*, 517–530.

Beinborn, L., Zesch, T., & Gurevych, I. (2015). Candidate evaluation strategies for improved difficulty prediction of language tests. *Proceedings of the Tenth Workshop on Innovative use of NLP for Building Educational Applications*, 1–11.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1–11.

Brizuela, A., & Montero-Rojas, E. (2014). Prediction of the difficulty level in a standardized reading comprehension test: contributions from cognitive psychology and psychometrics. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, *19*(2). https://doi.org/10.7203/relieve.19.2.3149

Davies, M. (2008). The Corpus of Contemporary American english (COCA). https://www.english-corpora.org/coca/.

Davies, M. (2011). Most frequent 100,000 word forms in English (based on data from the COCA corpus). https://www.wordfrequency.info/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.48550/ARXIV.1810.04805

Deza, M. M., & Deza, E. (2016). *Encyclopedia of Distances* (4th ed.). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-52844-0

Ferrara, S. (2022). Response demands of reading comprehension test items: A review of item difficulty modeling studies. *Educational Assessment*, *27*(1), 1–21. https://doi.org/10.1080/08957347.2022.2103135

Gunawan, D., Sembiring, C. A., & Budiman, M. A. (2018). The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. *Journal of Physics: Conference Series*, *978*, 012120. https://doi.org/10.1088/1742-6596/978/1/012120

Hvitfeldt, E., & Silge, J. (2021). *Supervised machine learning for text analysis in R*. Chapman; Hall/CRC. https://doi.org/10.1201/9781003093459

Chall, J. S., & Dale, E. (1995). *Readability revisited: the new Dale-Chall readability formula*. Brookline Books.

Chen, X., & Meurers, D. (2017). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. https://doi.org/10.1111/1467-9817.12121

Kincaid, J. P., Fishburne, R. P. J., Rogers, R. L., & Chissom, B. S. (1975). Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. *Institute for Simulation and Training*, *56*. https://stars.library.ucf.edu/istlibrary/56

Martinková, P., & Hladká, A. (2023). *Computational Aspects of Psychometric Methods: With R*. Chapman; Hall/CRC. https://doi.org/10.1201/9781003054313

McLaughlin, G. H. (1969). SMOG Grading: A New Readability Formula. *Journal of Reading*, *12*(8), 639–646.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space [arXiv: 1301.3781]. *arXiv:1301.3781 [cs]*. Retrieved December 30, 2021, from http://arxiv.org/abs/1301.3781

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. http://www.aclweb.org/anthology/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

Štěpánek, L., Dlouhá, J., & Martinková, P. (2023). Item Difficulty Prediction Using Item Text Features: Comparison of Predictive Performance across Machine-Learning Algorithms. *Mathematics*, *11*(19), 4104. https://doi.org/10.3390/math11194104

Tan, P.-N., Steinbach, M., & Kumar, V. (2005, May). *Introduction to Data Mining*. Pearson.

Tränkle, U., & Bailer, H. (1984). Kreuzvalidierung und Neuberechnung von Lesbarkeitsformeln für die deutsche Sprache. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *16*(3), 231–244.

Wilbur, W. J., & Sirotkin, K. (1992, February). The automatic identification of stop words. https://doi.org/10.1177/016555159201800106

Xia, P., Zhang, L., & Li, F. (2015). Learning Similarity with Cosine Similarity Ensemble. *Information Sciences*, *307*, 39–52. https://doi.org/10.1016/j.ins.2015.02.024

Zhou, K., Ethayarajh, K., Card, D., & Jurafsky, D. (2022). Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 401–423.